Title

hyphy: Deep Generative Conditional Posterior Mapping of Hydrodynamical Physics

Permalink

https://escholarship.org/uc/item/47w1w596

Journal

The Astrophysical Journal, 941(1)

ISSN

0004-637X

Authors

Horowitz, Benjamin
Dornfest, Max
Lukić, Zarija
et al.

Publication Date

2022-12-01

DOI

10.3847/1538-4357/ac9ea7

The Astrophysical Journal, 941:42 (11pp), 2022 December 10 https://doi.org/10.3847/1538-4357/ac9ea7

HYPHY: Deep Generative Conditional Posterior Mapping of Hydrodynamical Physics

Benjamin Horowitz[1,2], Max Dornfest[2,3], Zarija Lukić[2], and Peter Harrington[2]
[1] Department of Astronomy, Princeton University, Princeton, NJ 08544, USA; bhorowitz@princeton.edu
[2] Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA 94720, USA; Dornfest@berkeley.edu
[3] Department of Physics, University of California at Berkeley, Berkeley, CA 94720, USA

Received 2021 July 14; revised 2022 August 23; accepted 2022 October 24; published 2022 December 12

Abstract

Generating large-volume hydrodynamical simulations for cosmological observables is a computationally demanding task necessary for next-generation observations. In this work, we construct a novel fully convolutional variational autoencoder (VAE) to synthesize hydrodynamic fields conditioned on dark matter fields from *N*-body simulations. After training the model on a single hydrodynamical simulation, we are able to probabilistically map new dark-matter-only simulations to corresponding full hydrodynamical outputs. By sampling over the latent space of our VAE, we can generate posterior samples and study the variance of the mapping. We find that our reconstructed field provides an accurate representation of the target hydrodynamical fields as well as reasonable variance estimates. This approach has promise for the rapid generation of mocks as well as for implementation in a full inverse model of observed data.

Unified Astronomy Thesaurus concepts: Intergalactic medium (813); Extragalactic astronomy (506); Intergalactic gas (812); Convolutional neural networks (1938); Neural networks (1933)

1. Introduction

Understanding the large-scale structure of the universe requires simultaneous analysis of both the evolution of the underlying dark matter cosmic web and the complex hydrodynamics leading to the formation of biased tracers. Over the past thirty years, hydrodynamical simulations have become the standard tool to generate mock observable data that includes both of these effects (Evrard 1990; Cen 1992; Katz et al. 1996; Springel 2005). However, the power of these hydrodynamical simulations comes with significant computational cost, and the next generation of cosmological surveys will require unprecedented precision across a wide range of scales (e.g., Walther et al. 2021). In this regime, computing quantities like covariance matrices (which require large numbers of simulations) becomes an increasingly daunting task, so there is a clear need for approximate methods that can ease some of the computational burden.

In recent years, machine-learning techniques have emerged as promising surrogate models for complex hydrodynamics, as they can be used to rapidly generate hydrodynamic fields with remarkable perceptual and statistical fidelity. In Zamudio-Fernandez et al. (2019), the authors were able to generate realistic neutral hydrogen (H I) maps which reproduce the properties of hydrodynamical simulations over a range of scales. In Tröster et al. (2019), the authors used generative models to map from two-dimensional (2D) dark matter maps to thermal Sunyaev–Zeldovich (tSZ) maps. They were able to reproduce accurate tSZ summary statistics over a wide range of scales, given only the dark matter maps. Related work in Wadekar et al. (2020) used a more traditional feed-forward architecture, HInet, to paint neutral hydrogen in all three-dimensions (3D), but this architecture does not allow exploration of posterior properties and uncertainties.

Estimation of the uncertainty of a neural network's output is critical in order to propagate errors accurately for cosmological and cosmographical analysis. However, within the astronomical community there has been relatively little work in error analysis in the context of these neural-network-based surrogate models. A promising approach in the case of low-dimensional data is to make the network output be a multidimensional Gaussian distribution as opposed to a single-point estimation, i.e., a Gaussian mixture model (see, e.g., Tsang & Schultz 2019). However, for high-dimensional outputs this approach would have difficulty capturing the full covariance in a memory-efficient way.

In this work, we will instead structure the latent space of a conditional variational autoencoder (C-VAE) to learn the uncertainty in the mapping from a dark matter map to the hydrodynamical quantities. The general structure of our network is inspired by style-transfer machine-learning literature (Johnson et al. 2016; Esser et al. 2018), where the latent space of the C-VAE is used to capture stylistic characteristics of the mapping. A similar network has recently been used to generate a realistic distribution of galaxy images (Lanusse et al. 2021).

As we expect hydrodynamic quantities to be quasi-local, we constrain our model to maintain this property by restricting the spatial field of view of the input and use convolutional layers of different sizes in order to capture information across a range of scales. When transforming to redshift space, as is needed to match observables, this locality is broken, so rather than directly modeling observable quantities we instead focus on reconstructing the underlying (real space) baryon density, temperature, and velocity fields. Then, estimates for the target observables can trivially be computed using existing analytical tools, which also allows flexibility in modeling physical details "orthogonal" to the dark matter and hydrodynamics relationship (e.g., atomic species ratios, ionization rates, etc). For this work we will focus on Lyα flux for Lyα forest cosmology measurements, but our model outputs are generic and could be applied to generate other target observables.

This work is a companion work to Harrington et al. (2022), which used a deterministic network to perform this mapping.

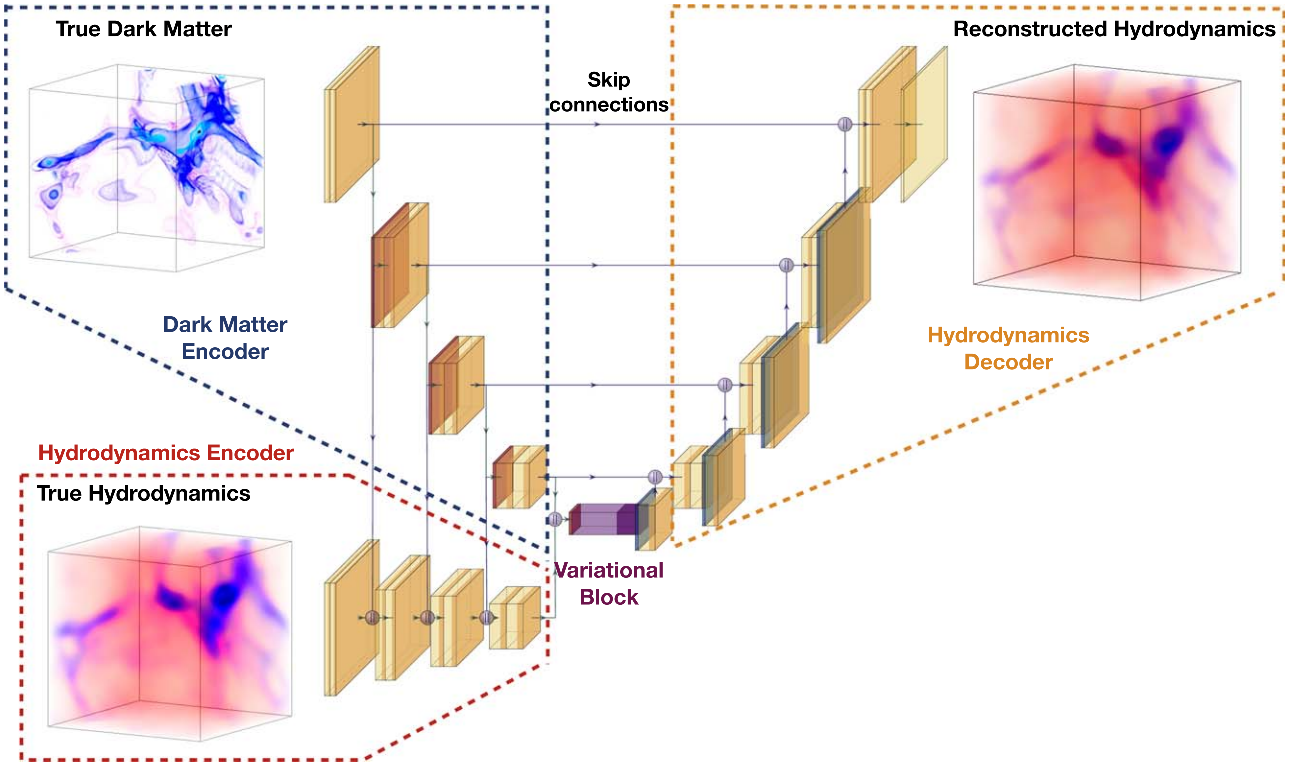

Figure 1. Schematic showing the flow of our U-Net C-VAE architecture. Our model can be viewed as four interconnected parts: an encoder for the dark matter fields, an encoder for the hydrodynamical fields, a variational block, and a decoder. The network is constructed such that after training the hydrodynamical block can be removed and the latent space sampled from a unit Gaussian to generate the corresponding hydrodynamical field. Plotted is only the dark matter density and the baryon temperature, but the model also is fed as an input the dark matter velocity and outputs the baryon density and line-of-sight velocity. The term "dark matter encoder" is used as the model can be viewed as encoding the dark matter structure over the course of training, but we do not explicitly train a model to decode this field. The full network architecture can be found online at https://github.com/bhorowitz/HyPhy-public.

While this approach successfully recovers the key Lyα observables for next-generation cosmological measurements, it has difficulty in capturing stochastic processes such as shocked regions. As the approach and goal of these works is quite different we present them as two separate papers.

The paper is organized as follows. In Section 2 we briefly describe NYX and the simulation data set used and then describe our neural network architecture. We present our results in Section 2, first reviewing maximum a posterior (MAP) performance and posterior accuracy. We conclude Section 4 with our ongoing work and future areas of interest to the community.

2. Methodology

2.1. Hydrodynamical Simulations

We choose to obtain simulation data from NYX, a massively parallel multiphysics code, because it was developed for simulations of the intergalactic medium (IGM) and has been used for many recent IGM studies (Davies et al. 2020; Horowitz et al. 2019; Walther et al. 2019), and is capable of modeling dark matter and hydrodynamic evolution in great detail. The NYX code (Almgren et al. 2013) follows the evolution of dark matter modeled as self-gravitating Lagrangian particles, while baryons are modeled as an ideal gas on a set of rectangular Cartesian grids. The Eulerian gas dynamics equations are solved using a second-order accurate piecewise parabolic method to accurately capture shocks. Besides solving for gravity and the Euler equations, we also include the main physical processes relevant for the Lyα forest. We consider the chemistry of the gas as having a primordial composition of hydrogen and helium, include inverse Compton cooling of 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 type of simulation is used as a forward model in virtually any recent inference work using a Lyα power spectrum (Boera et al. 2019; Walther et al. 2019, 2021; Palanque-Delabrouille et al. 2020; Rogers & Peiris 2021). Simulations of this kind neglect the effects of inhomogeneous reionization, which produces temperature and UV background fluctuations on large scales, especially at redshifts higher than those studied in this work ($z \gtrsim 4$).

In this work we used simulations of a standard Lambda cold dark matter (ΛCDM) cosmological model, consistent with the latest cosmological constraints from the cosmic microwave background (CMB; Planck Collaboration et al. 2020): $\Omega_{\rm m} = 0.31$, $\Omega_\Lambda = 0.69$, $\Omega_b = 0.0487$, $h = 0.675$, $\sigma_8 = 0.82$, and $n_s = 0.965$. For the hydrogen and helium mass abundances we adopted values consistent with CMB observations and Big Bang nucleosynthesis (Coc et al. 2013): $X_p = 0.76$ and $Y_p = 0.24$. The box size of simulations is 20 h^{-1}Mpc, with $N = 1024^3$ particles and grid cells, resulting in a resolution of $\sim$20 h^{-1} kpc, fulfilling convergence criteria for percent-level accurate Lyα quantities (Lukić et al. 2015).

In addition to hydrodynamical simulations, we have also produced N-body simulations starting with the same initial conditions. These neglect all other forces but gravity, and all matter is considered collisionless, although baryonic effects are imprinted in the initial power spectrum for the total matter. Baryons have only a minor effect on dark matter clustering in the regime relevant for the Lyα forest, but we have nevertheless produced these "companion" simulations to maintain maximum reality in our modeling when we train to infer hydrodynamical quantities from a N-body run. We do not want our model to learn about the baryonic field through back reaction on the dark matter field, so these companion simulations eliminate this risk. Throughout this work we use one set of hydro and matching N-body simulation for training and a different set for validation purposes. The two have matching power spectra and differ only in the random phases of the Gaussian random field in the initial conditions.

Since our work is focused on predictions for Lyα cosmological analysis, we are mostly interested in the baryonic quantities which go into predictions of Lyα flux, namely the baryon density, temperature, and line-of-sight velocity (for redshift space distortions). These are the quantities we predict based on the dark matter density and dark matter line-of-sight velocity. One could further expand this to predict all velocity fields, but we found in practice this significantly increased the computational cost for training in both memory and GPU time.

2.2. Conditional Variational Autoencoder Model

As neural networks are becoming an increasingly well-known tool in astrophysics and cosmology, we briefly summarize our model[4] and highlight how it differs from others in the literature.

We use a C-VAE architecture (Sohn et al. 2015; Gu et al. 2018) to study the posterior of hydrodynamical quantities, τ, given the dark matter field, δ. The core concept of a VAE network (Kingma & Welling 2013; Kingma et al. 2014) is that the neural network learns the probability space described by the training sample by marginalizing over a (usually bottlenecked) set of latent space parameters, Z. In this case we are interested in constraining the outputs of our model to those corresponding to a given dark matter realization, which we enforce by using the dark matter realization as a "condition": a field given to the network both during the "encoding" step (where the fields are mapped to the latent space) and the "decoding" step (where the latent space is mapped to a 3D field). For finding a given hydrodynamical realization, we are interested in calculating the following quantity:

$$P(\tau|\delta) = \int p(\tau|Z,\,\delta)p(Z|\delta)dZ, \tag{1}$$

where $p(\tau|Z,\,\delta)$ is the generator network. In order to train the network, we also need to define an overlapping encoder network, $q(Z|\tau,\,\delta)$. We can generalize the standard evidence lower bound straightforwardly to include this conditioning variable:

$$\begin{aligned}\log P(\tau|\delta) &= \log \int p(\tau|Z,\,\delta)p(Z|\delta)dZ \geqslant \mathbb{E}_q\left[\log \frac{p(\tau,\,\delta|Z)}{q(Z|\tau,\,\delta)}\right]\\ &= \mathbb{E}_q\left[\log \frac{p(\tau|\delta,\,Z)p(Z|\delta)}{q(Z|\tau,\,\delta)}\right],\end{aligned} \tag{2}$$

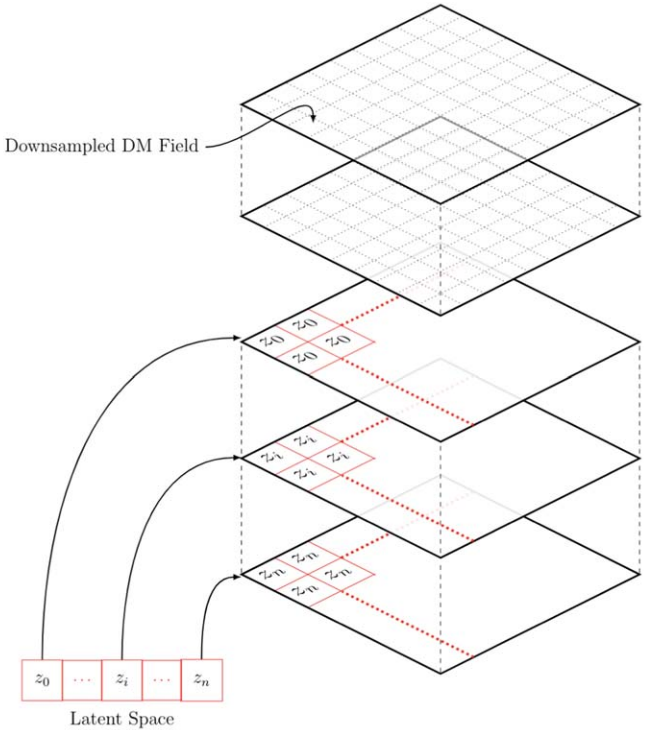

Figure 2. Diagram showing the broadcasting of the latent space samples into the same dimensionality as the downsampled dark matter field. This stack will comprise the filters passed to the upsampling convolutional layers in the hydrodynamics decoder. This allows changing input dark matter map sizes during generation while keeping the same dimensionality of the latent space and avoiding the need for dense layers. The presence of dense layers during generation would implicitly fix the box size and not allow a fully convolutional structure. Note that in our model each filter is three-dimensional, not two-dimensional as shown here.

where $\mathbb{E}_q$ is the expectation value over q. In standard style-transfer implementations, $p(Z|\delta)$ plays an important role as a prior over the latent space parameters (i.e., Esser et al. 2018). In our case, we will not be imposing any direct interpretation to our latent space parameters, and we will find that this prior distribution, $p(Z|\delta)$, will be pulled to a unit-normal Gaussian due to the loss term.

We again follow the standard derivation for C-VAE networks and model the probability distributions, $p(\tau|\delta,\,Z)$ and $q(Z|\delta,\,\tau)$, as neural networks with associated free parameters, G_θ and F_ϕ, respectively. In Figure 1, G_θ corresponds to the combination of the dark matter encoder and hydrodynamics decoder, while F_ϕ corresponds to the combination of dark matter encoder and hydrodynamics encoder. These overlapping neural networks' weights, θ and ϕ, are trained jointly using the standard Adam optimizer (Kingma & Ba 2014) under the associated loss:

$$\begin{aligned}\mathcal{L}(\delta,\,\theta,\,\phi) = &-\mathrm{KL}(q_\phi(Z|\delta,\,\tau)||p_\theta(Z|\tau))\\ &+ \mathbb{E}_{q_\phi(Z|\delta,\tau)}[\log p_\theta(\tau|Z,\,\delta)],\end{aligned} \tag{3}$$

where KL is the Kullback–Leibler divergence (Kullback 1997) to compare distributions. Assuming we treat the generator network, G_θ, as deterministic in δ and Z, we can simplify the second term with our chosen reconstruction loss. There are many possible choices for this loss function, including $L1$, $L2$, perceptual loss, or an adversarial loss. In our case we choose

[4] Our model is implemented in TENSORFLOW, with sample data and a saved model available at https://github.com/bhorowitz/HyPhy-public.

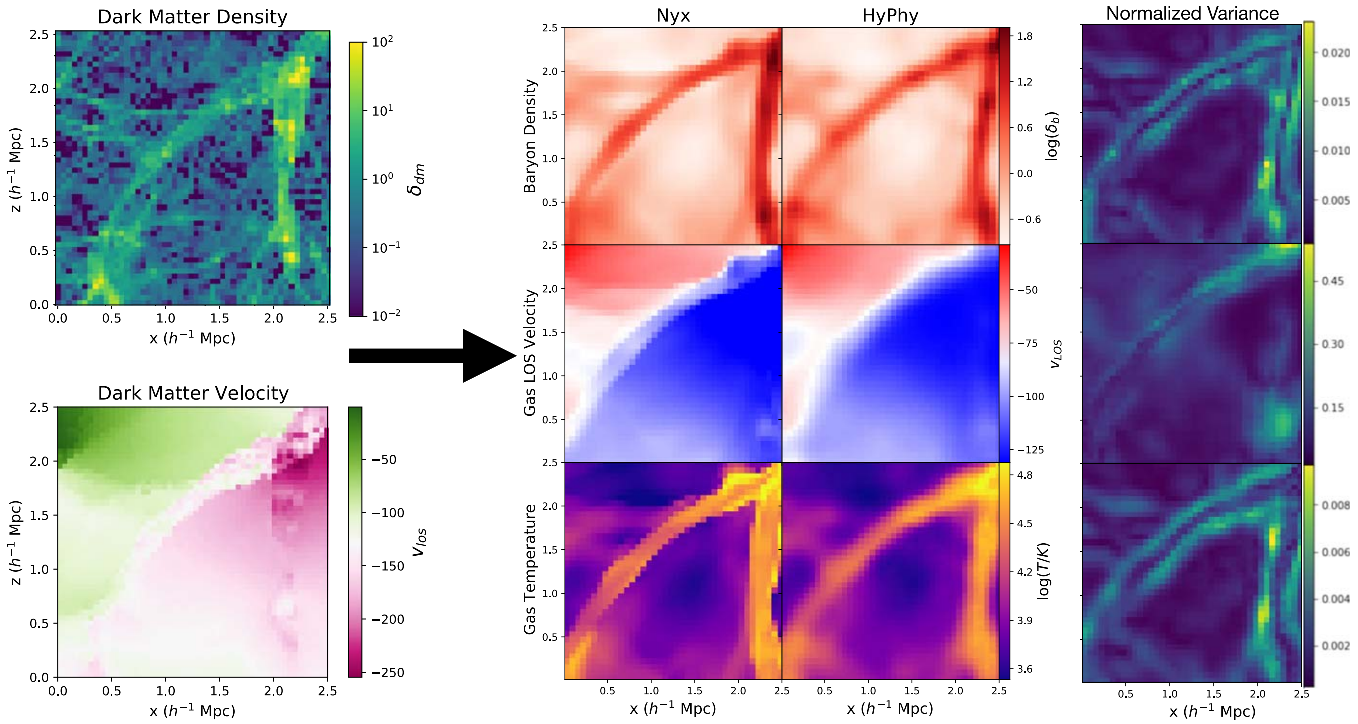

Figure 3. A typical example single slice of the HYPHY mapping is shown. On the left are the test input dark-matter-only maps and in the center are the hydrodynamical fields (baryon density, velocity, temperature) are compared. There is strong qualitative agreement, with the network accurately learning various characteristics of the hydrodynamical fields including the variable baryon pressure smoothing and thermal properties. On the far right the networks' estimated variance is shown, calculated from 1000 samples over the latent space, normalized by the mean value of the field.

the standard $L1$ loss term, i.e.,

$$\mathcal{L}(\delta, \theta, \phi) = -\mathrm{KL}(q_\phi(Z|\delta, \tau)||p_\theta(Z|\tau)) + ||\tau - G_\theta(\delta, Z)||_1, \tag{4}$$

where τ is the simulated true field corresponding to δ. The KL divergence term can also be further simplified by expressing our latent space image, $q_\phi(Z|\delta, \tau)$, as a multidimensional Gaussian with diagonal covariance using the reparametrization trick (Kingma & Welling 2013) and our target distribution as a unit normal Gaussian:

$$\begin{aligned}\mathcal{L}(\delta, \theta, \phi) = -\mathrm{KL}(\mathcal{N}(\mu(\tau, \delta), \sigma(\tau, \delta))||\mathcal{N}(0, 1)) \\ + ||\tau - G_\theta(\delta, Z)||_1,\end{aligned} \tag{5}$$

where μ and σ are outputs of our encoding network F_ϕ, and correspond to the means and standard deviations of our latent space distributions from which we sample. The variations of the hydrodynamical model are captured by properties of the dark matter field at a variety of scales. We are therefore motivated by examples of image segmentation from machine-learning literature to use an altered U-Net structure (Ronneberger et al. 2015) for our network. We summarize this structure in Figure 1. The most notable element in the structure of this network is the skip connections across the bottleneck mapping the dark matter field from the encoder to decoder side. This is designed to maximize the possible information extracted from the dark matter field, in a computationally expedient fashion, as well as to minimize the dark-matter-dependent structure in the latent space. In particular, since we draw samples from the latent space that are fully representative of uncertainty in the mapping, we want the conditional probability of $p(Z|\delta)$ to be as close to a uniform Gaussian, $\mathcal{N}(0, 1)$, as possible. Structure in the latent space will result in a biased posterior sample distribution. In abstract we could implement various techniques (such as an autoregressive flow, as in Lanusse et al. 2021), to ensure that our latent space does not encode information found in the dark matter structure alone (i.e., specific cosmic environments are not clustered in a certain region of latent space), but in practice we find that there is little dependency on matter properties in the latent space.

In addition, we have vertical skip connections during training to pass the downsampled dark matter field to the downsampled baryon field at each step of training. This allows the network to learn the expressive relationships between the dark matter field and the baryon field at each step of training in order to have the representative latent space during the bottlenecking step. In full generality one could simply concatenate an additional copy of the dark matter field with the baryon field in the hydrodynamics encoder, but this would significantly increase the memory requirements for the network.

2.3. Fully Convolutional Architecture

A critical feature of our architecture is for it to be fully convolutional, thereby allowing inputs of any size during inference while being trained on smaller volumes. Traditionally, VAEs utilize dense layers to upsample the drawn latent space variable, which are then possibly followed by transposed convolutions (or other upsampling convolutions) to reach the

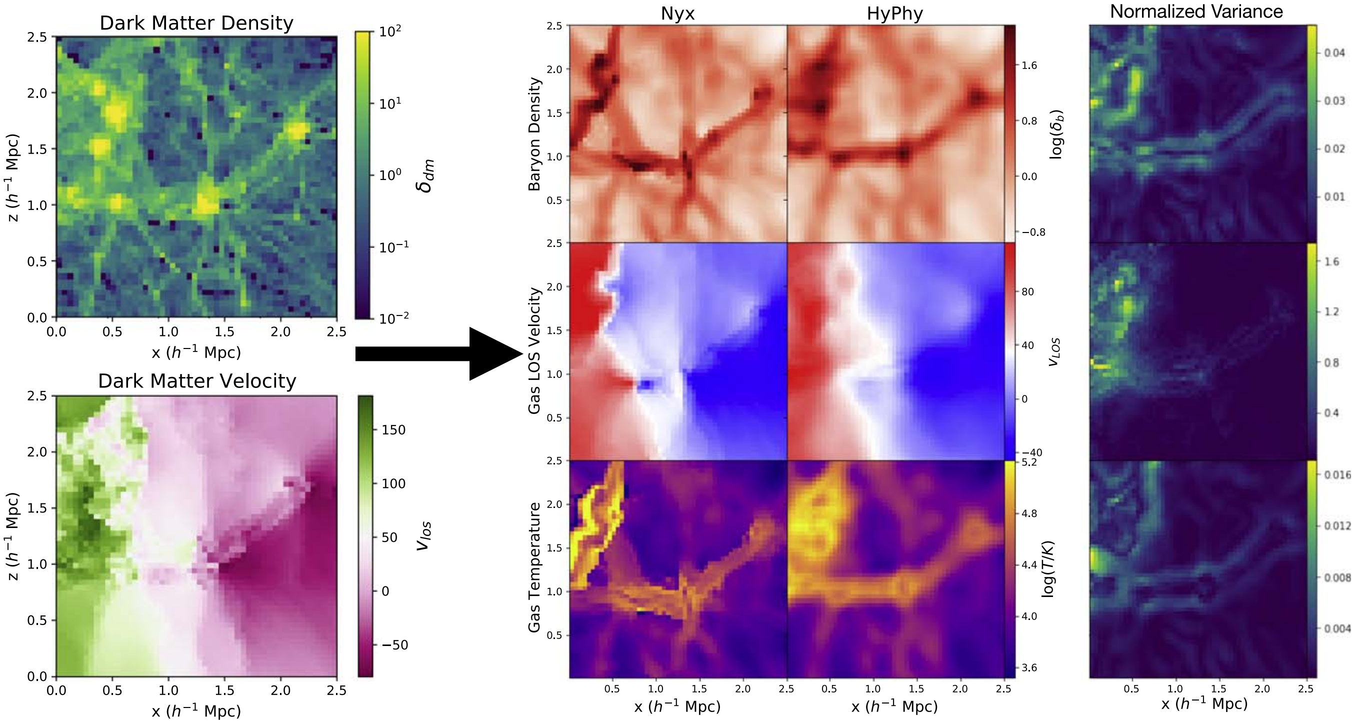

Figure 4. Same as Figure 3, but showing one of the worse reconstructed test boxes. The characteristics of moderately large shocks are difficult for our network to learn, resulting in a less sharp feature. However, this uncertainty in the reconstruction is captured when sampling over latent space, as we see significant variance in the shocked region. If we look at the individual posterior samples this is even more visually apparent; see Figure 9.

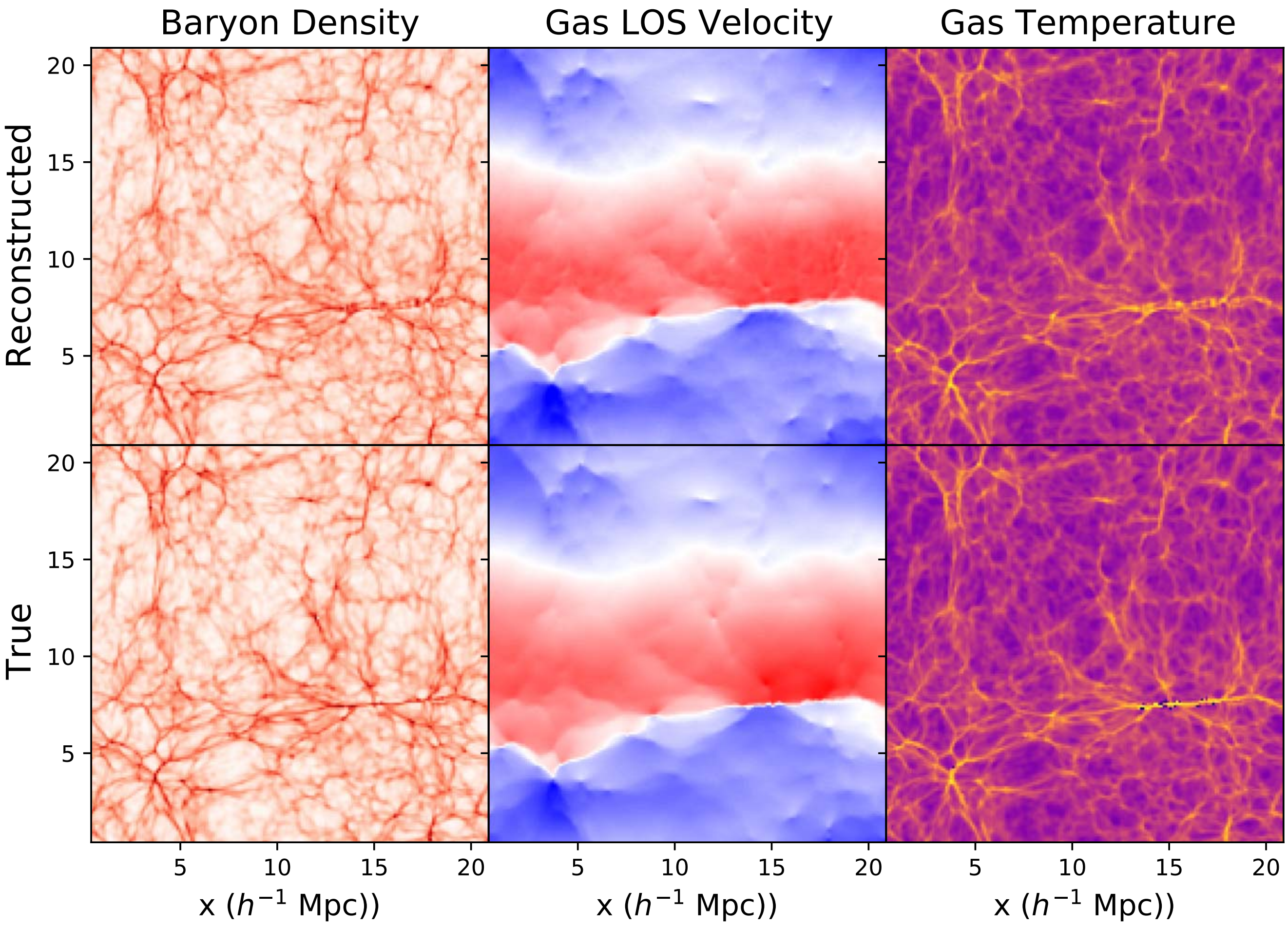

Figure 5. HYPHY run on a box significantly larger than the training volumes. The fully convolutional nature of the network allows the network to be run on any box that is a multiple of a base dimension (currently 8 pixels) and has the same spatial resolution. There are no obvious artifacts caused by the increased image size.

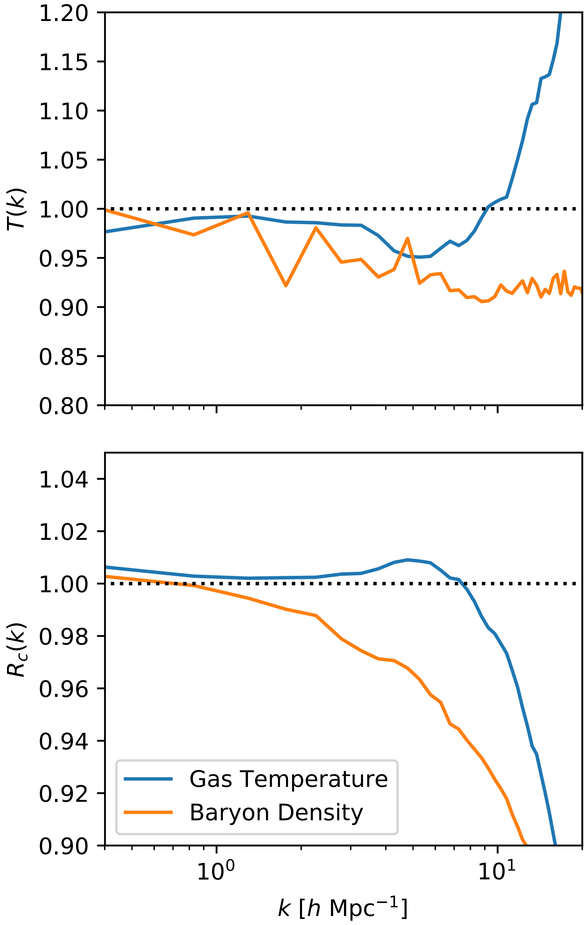

Figure 6. Top: transfer function, defined as the ratio between the large-volume HYPHY run with the simulated NYX. Bottom: correlation coefficient between the two boxes, defined as $R_c(k) = P_{\mathrm{nyx,hyphy}} / \sqrt{P_{\mathrm{nyx}} P_{\mathrm{hyphy}}}$, where $P_{\mathrm{nyx,hyphy}}$ is the cross-power between the two fields.

desired output size. This design restricts the input image size to always be of the same dimension as the training set. To avoid this constraint, we broadcast the latent space parameters into a feature map whose dimensions match that of the lowest level in the U-Net, then upsample from there as is standard (see Figure 2). Since the upsampling convolutional layers are inherently local (as determined by their kernel size) we maintain the desired locality of our network, while still allowing every element of the output to "see" the full latent space.

While generated volumes can be of any dimension, the training volumes are fixed in size due to the need for dense layers in the encoder to predict the latent space distributional parameters. These dense layers do not appear in the generation network as during generation the latent space is sampled, not predicted from the input fields. However, training data can easily be segmented to the proper size, and the choice of the crop size is a problem-dependent hyperparameter to be tuned.

Note that while one can apply the network to arbitrarily sized boxes, even if trained to perfect accuracy, its predictions will be limited by the amount that small-box simulations match large-volume simulations. Training on boxes of limited volume means long-distance correlations not captured in the dark matter distribution (e.g., a spatially fluctuating UV background) would not be well reproduced with this architecture and would require additional considerations. For the purpose of this work, we use training boxes of approximately 4 h^{-1} Mpc.

Since neither each individual training box nor standard convolutional layer implementations are periodic, we must minimize edge effects by restricting our loss function to only compare the central region of our training sample. In this work, we train on boxes with 64 voxel side length, of which we ignore 10 voxels on each side in order to avoid dealing with edge effects.

In order to better utilize our limited training data, we use the symmetries of our system to randomly augment training samples by performing reflections and rotations of our box over each of the three spatial dimensions. In addition, our training boxes are overlapping in space, providing some knowledge of the translational symmetry of the underlying physics model.

2.4. Redshift Information

For application to real Lyα forest data, a key aspect of the mapping is to include the redshift dependence of the mapping as this dependence is used for cosmological constraints (i.e., Chabanier et al. 2019). To include this in our model, we condition on a redshift field of the same size as our training volume concatenated onto the input dark matter field (i.e., every pixel has an associated redshift). This allows us to vary the redshift over the box (i.e., to generate light cones), as well as train the same model to work across cosmic time. We train our model using snapshots from the same simulation at $z = 2.4$, $z = 3.0$, and $z = 4.0$, as this range dominates the cosmological Lyα forest signal (Walther et al. 2021). The specific number and range of bins to use depends on the specific application; here we focus on qualitative features of the resulting map.

3. Results

We apply our trained network to a separate simulation (not used for training), focusing on the central redshift of $z = 3.0$, which requires only changing the latent space projection dimension while maintaining the same network weights. We first show the results for the MAP point predicted by the network, both in the base hydrodynamical quantities and in terms of derived Lyα flux. We then discuss the posterior properties of a representative sample distribution in Section 3.2.

The MAP should correspond to the highest maximum of the distribution $P(\mathbf{Z})$, which by construction will be at $\mathbf{Z} = \mathbf{0}$ for a fully trained model. A useful starting point for our analysis will be to see how this point in the posterior space behaves to judge the quality of our reconstruction for test boxes with the same dimension as our training set. Note that we expect these reconstructions to be slightly smoothed versus generic samples.

We show test boxes both without and with significant shocked regions in Figures 3 and 4, respectively. In low to medium density regions, we recover excellent qualitative agreement across a range of scales for baryon density, baryon velocity, and temperature. For the shocked example (Figure 9), the HYPHY-recovered temperature field has a significantly less prominent shocked region with a much smoother boundary; this is a common phenomenon in VAE-type networks (Khan et al. 2018).

We also show the model-predicted variance by examining 1000 samples drawn from $\mathcal{N}(0, 1)$, qualitatively showing that the regions of highest variance are where there is the strongest disagreement between HYPHY output and the simulated true

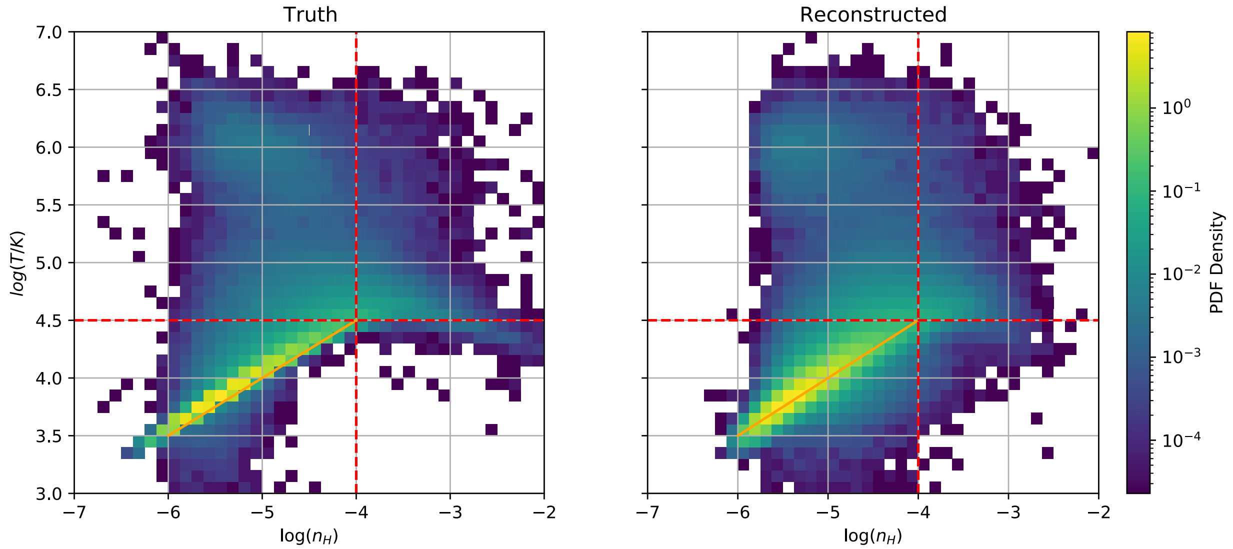

Figure 7. Two-dimensional histogram of temperature and hydrogen density relation in truth compared to the reconstruction. The orange line indicates a best-fit power-law solution, $T = T_0 n_{\mathrm{H}}^{\gamma}$, across the range of $10^{-4} < n_{\mathrm{H}} < 10^{-6}$. Red dashed lines indicate the classification of the gas content of the universe into wCGM, WHIM, diffuse gas, and halo gas. We find excellent agreement in terms of the mean relation, as well as the WHIM component. However, the cool highest-density regions, corresponding to centers of cluster regions (i.e., CGM regions), are not well reconstructed.

field. In particular, regions near the boundaries of structures (like filaments) have significant variance as do those with significant astrophysical shocks. We discuss these properties more in Section 3.2.

3.1. Large-volume Statistics

Next, we apply HYPHY to the entire test N-body simulation at once. To match the resolution of our training sample, we uniformly downsample the volume by a factor of 2. We show the reconstructed resulting baryon density, velocity, and temperature in Figure 5. In addition, we calculate and compare a number of the statistics between the simulated truth and the HYPHY-generated version in Figure 6.

3.1.1. Gas-phase Physics

A well-studied aspect of cosmological hydrodynamical simulations is the relation between the gas density and gas temperature (Gunn & Peterson 1965; Sorini et al. 2016). We show our reconstructed relation in log space in Figure 7. Following Ursino et al. (2010), Martizzi et al. (2019), and Galárraga-Espinosa et al. (2021), this plot can be viewed as a phase-space distribution between warm hot intergalactic medium (WHIM), warm circumgalactic medium (wCGM), diffuse IGM, halo gas, and "hot gas." Halo gas consists of relatively "cool" gas including the interstellar medium within galaxies, as well as more diffuse gas found in between galaxies within halos. Similarly, wCGM is found in dense environments but has been significantly heated via shocks or feedback processes near galaxies. Diffuse IGM and WHIM are found in less dense environments, such as regions surrounding filaments. The WHIM component is of significant interest due to the "missing baryon" problem (Fukugita et al. 1998). The last component, consisting of gas at any density at a temperature in excess of 10^7 K and generally associated with massive shocks, is a vanishingly small percentage of the test volume (68 pixels out of 512^3 total pixels) and is grouped with wCGM or WHIM depending on density. In addition, some studies (e.g., Martizzi et al. 2019) separate star-forming gas, with densities $\log(n_{\mathrm{H}}) > -1.0$, as a separate phase. Since NYX does not model star formation we do not include this phase in our analysis.

In Figure 7, we summarize the recovered volume fractions versus the simulated truth. We find excellent recovery of the diffuse IGM and WHIM, with slightly worse performance of the wCGM component. Halo gas, constituting a tiny fraction of the total volume, is very difficult to recover accurately, with HYPHY finding a factor 19× more halo gas. However, it is important to note the vast majority of this excess is on the border of the diffuse IGM component in phase space and can probably be better described as overestimated IGM temperature as opposed to misattributed halo gas.

A key property of interest to the IGM is the power-law relation between density and temperature, which can be related to statistics of the Lyα forest (Hui & Gnedin 1997). Following the procedure in Sorini et al. (2016), we identify the gas around two bins centered at $\log(\Delta_{b,0}) = -1$ and $\log(\Delta_{b,1}) = 0$, with a width of 5% around the central value. We calculate the median temperature of the corresponding particles in each bin and use those two points to determine the power-law relation $T = T_0 \Delta_b^{\gamma}$, where Δ_b is the hydrogen density. For NYX, we find $(T_0, \gamma) = (10^{4.08}\mathrm{K}, 1.53)$ and for the HYPHY reconstruction we find $(T_0, \gamma) = (10^{4.08}\mathrm{K}, 1.51)$.

3.2. Posterior Exploration

The main motivation of our architecture is to allow accurate, unbiased posterior sampling of our hydrodynamical quantity through Gaussian sampling of our latent space variable. In this subsection we explore how accurate is this mapping.

This question is not straightforward as it is computationally difficult to calculate the true posterior for the target mapping, which would essentially require running potentially millions of additional hydrodynamical simulations to explore all possible evolutions resulting in the same binned density and velocity fields due to variations of particle positions on the sub-binned scale. With this in mind, we are left to check if the samples drawn have the correct statistical properties of an a posteriori

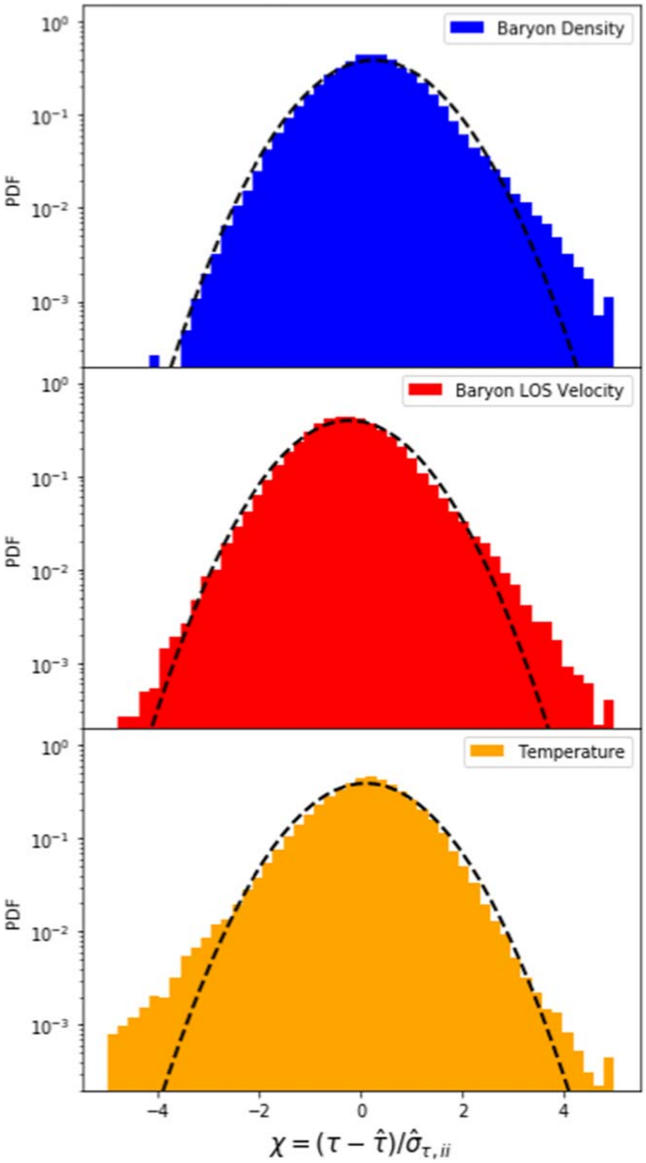

Figure 8. Distributions of the χ value for each of our three reconstructions in relation to a single sample. Also plotted is a unit normal Gaussian distribution, showing close agreement. Excess values along the tails indicate model failure, either to estimate proper variance or strong residual biases, or potentially the importance of off-diagonal covariance terms not captured in this analysis. We find these value points constitute a very small volume fraction ($\lesssim$0.1%). The field with the most variance that is not captured by our model is the temperature field, which is also the most numerically complex to calculate in the original hydrodynamical simulation due to shock physics.

sample, as well as examine its qualitative behaviors. We draw 1000 random samples from a unit Gaussian distribution in latent space and predict the hydrodynamical quantities associated with a test dark matter field. To test whether or not our variance estimates are accurate, we plot the χ values for a test sample in Figure 8. For a true posterior we would expect this distribution to be approximately unit Gaussian. Schematically, we would expect that points that diverge significantly from the true value of a given hydrodynamical quantity at a point would have high predicted variability at that point over many samples.

In particular, we can use the samples to construct a covariance matrix to use to test the statistical significance of deviations away from the MAP solution using the standard chi-squared formula,

$$\chi^2 = (\boldsymbol{\tau} - \hat{\boldsymbol{\tau}})^T \boldsymbol{C}^{-1}(\boldsymbol{\tau} - \hat{\boldsymbol{\tau}}) \rightarrow (\boldsymbol{\tau} - \hat{\boldsymbol{\tau}})^2/\sigma_\tau^2, \tag{6}$$

where $\hat{\tau}$ is the estimate hydrodynamical quantity in a single sample and τ is the simulated true value. In the last arrow we assume a diagonal covariance for implementation/memory reasons. If each sample is truly independent and represents the posterior, we expect that the corresponding χ^2 values from the ensemble should be Gaussian distributed with zero bias and a standard deviation of one. This is arguably a necessary, but not sufficient, condition for the samples to represent a maximum a posteriori sample. We show this distribution in Figure 8. Outlier samples are caused by a combination of model failures in rare environments without sufficient training data and by the limitations of our diagonal approximation for our χ statistic.

An additional property desired of the posterior is for the samples to capture the qualitative uncertainty of shocked regions when sampled over latent space. We show one example of such a region in Figure 9. We expect dense regions at nodes of the cosmic web to have the largest hydrodynamical effects, while those in underdense environments away from cosmic structures should follow roughly power-law distributions without uncertainty. In the samples in Figure 9, one sees high variability of the specifics of the shocked region, indicating the model is learning to account for hydrodynamical uncertainty. Again, it is difficult to formulate a rigorous method to test whether this variability is the "true" uncertainty without running a large suite of hydrodynamical simulations.

3.3. Predicted Lyα Forest

The Lyα forest arises from the scattering of photons at the characteristic rest-frame Lyα frequency along their path from a background source, generally either a quasar or galaxy, to the observer. The fraction of the transmitted flux is given by $F = \exp(-\tau)$, where τ is the optical depth of the intervening gas. The optical depth in redshift space at a given velocity coordinate, u, along the line of sight is given by

$$\tau(u) = \int du' \frac{\lambda_{\mathrm{Ly}\alpha}\sigma\, n_{\mathrm{H\,I}}(\boldsymbol{u}')}{\mathrm{H}(z) b(\boldsymbol{u}')} \times \exp\left[-\frac{(u - (u' + u_{\mathrm{pec}}(u')))^2}{b(\boldsymbol{u}')^2}\right], \tag{7}$$

where u' is the component of the Hubble flow velocity field $\boldsymbol{u}'$ along the line of sight over which the integral is calculated and $b(\boldsymbol{u}') = \sqrt{2k_B T(\boldsymbol{u}')/m_p}$ is the thermal broadening of the absorption feature. With the output of the HYPHY model, we have all the components necessary to calculate the predicted Lyα forest statistics. We run the GIMLET (Friesen et al. 2016) library to numerically calculate Equation (7) on both the original simulation and our predicted output, using a mapping from hydrogen fraction to neutral hydrogen from Rahmati et al. (2013). Flux distributions are calculated along each of three axes and then averaged out. In Figure 10, we show the resulting flux distribution, and in Figure 11 we show the error on the reconstructed 1D power spectra.

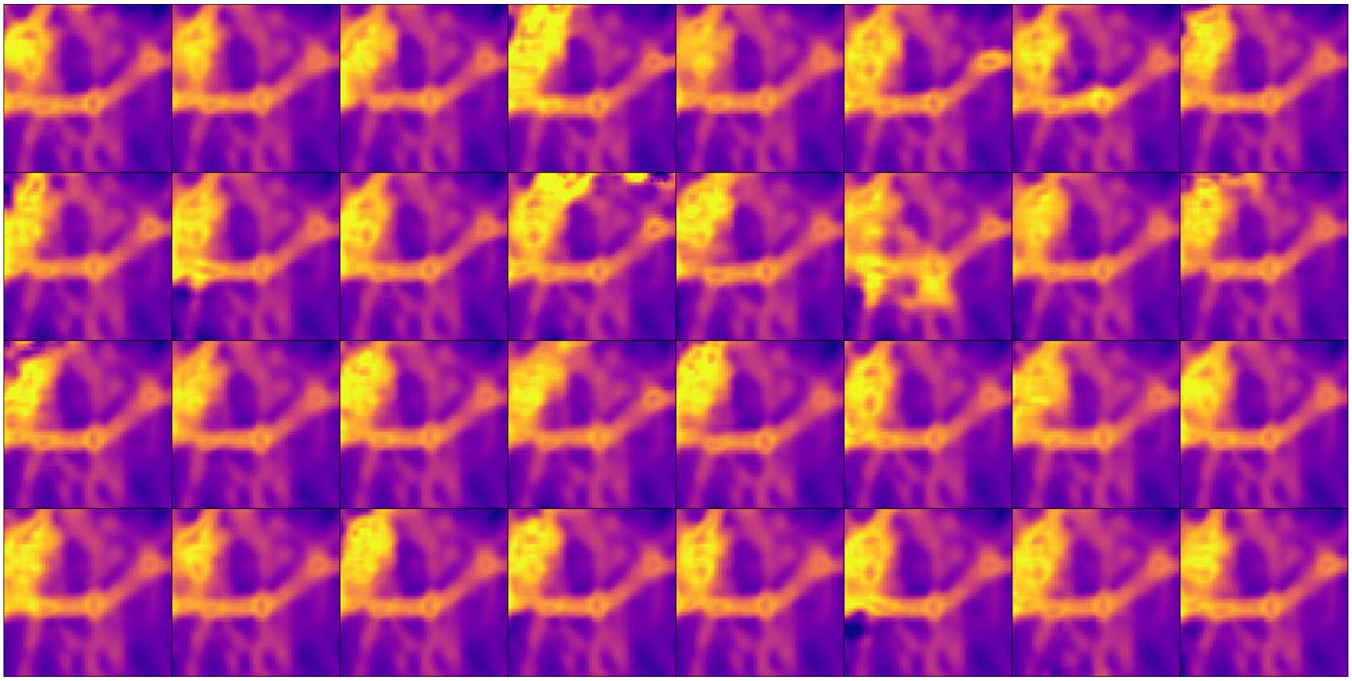

Figure 9. Postage stamp posterior samples of the temperature field from Figure 4. There is high variation in the shocked region, showing the hydrodynamical uncertainty the model has learned, while the areas outside the shocked region are stationary (with some occasional edge effects).

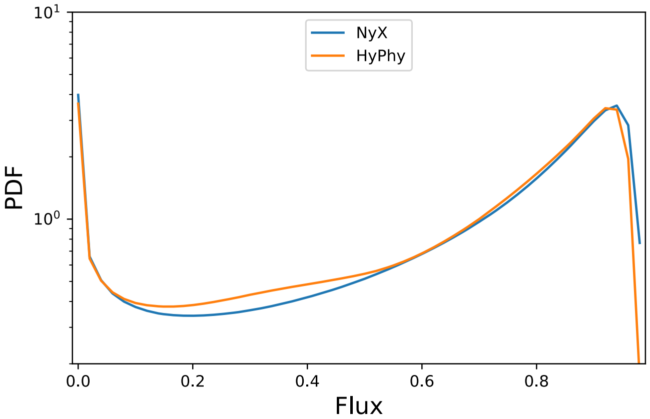

Figure 10. Performance of HYPHY in terms of the Lyα forest flux in redshift space using Equation (7) as implemented in GIMLET. We show the resulting probability density function of Lyα flux, showing substantial biases in the reconstructed flux at $F \sim 0.25$, but good agreement at the extremes.

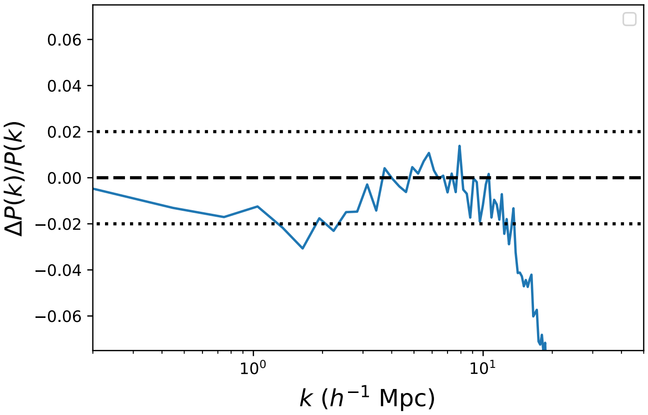

Figure 11. Performance of HYPHY in terms of the Lyα forest power spectrum in redshift space using Equation (7) as implemented in GIMLET. We show the resulting power spectra of Lyα flux, finding good agreement across a range of scales up to $k \sim 14\ h^{-1}$ Mpc.

4. Conclusion

In this work, we have provided a flexible generic mapping from dark matter fields to hydrodynamical quantities, in particular the gas temperature, density, and velocity. We demonstrated that this mapping provides accurate reconstructions across a wide range of scales and captures a number of the statistical properties of the underlying truth fields. In addition to constructing this mapping, the underlying model provides posterior samples which are consistent with the underlying dark matter field. While it is computationally infeasible to analytically calculate the true posterior in this case, we show that our posterior samples are a valid posterior through their variance properties.

It is important to note that we were able to construct these mappings despite only training on a single NYX-simulated run, using data at $z = 2.4$, $z = 3.0$, and $z = 4.0$. This was possible due to the various data augmentations used, which exploit the symmetry of the hydrodynamical physics as well as a regularizing effect of the underlying C-VAE architecture (Kamyab et al. 2019). Despite this relatively small training volume, we are still able to capture the qualitative effects and associated uncertainties in shocked regions.

We do find some cosmic environments where there are constant residual biases which, despite significant testing, we were unable to reduce completely. While we recover the statistical properties of the WHIM, wCGM, and diffuse IGM very well, recovering the hot halo gas contribution is significantly more difficult. We find that our model systematically underestimates the density in these regions while overestimating its overall volume fraction, resulting in an incorrect estimate for the phase-space distribution. Gas in this region accounts for a very small small percent of the total volume, $\sim 4 \times 10^{-6}$%, likely resulting in HYPHY having difficulty capturing its properties. Due to the high mass of these regions, they have an outsized effect on summary statistics such as the comparative transfer functions and cross-correlation measures, as shown in Figure 6.

However, for Lyα forest analysis, these wCGM regions do not have a noticeable effect on the forests' statistical properties so we do not focus on optimizing this aspect. Going beyond this work, one could perform various data augmentations to increase its importance in the loss to result in better model performance in this region. Approaches to dealing with such unbalanced data have been well studied in the machine-learning literature (Wang et al. 2016), including oversampling the minority data (Khoshgoftaar et al. 2007) and using generative adversarial networks to create synthetic minority data (Kiyoiti dos Santos Tanaka & Aranha 2019).

One of the most compelling applications of the HYPHY network is in forward modeling, where the underlying density field is reconstructed from observed data through an optimization process. For this application, the HYPHY network could replace analytical or semianalytical approximations, such as the fluctuating Gunn–Peterson approximation for the Lyα forest (e.g., in TARDIS; Horowitz et al. 2019, 2021). Our model is fully differentiable, allowing propagation through a model using first- or second-order methods. This is similar in spirit to work done in Modi et al. (2018), where a neural network was used to paint halo fields onto forward-modeling dark matter density. However, this assumed a deterministic mapping from dark matter to galaxy light, while in HYPHY this mapping is controlled by a latent space. In this approach, hydrodynamical uncertainties could be marginalized out via sampling of the latent space during optimization, or jointly optimized for and then marginalized out via variational methods.

In our loss function, we are implicitly assuming our latent space posterior can be well approximated by the dimensionality of our multivariate Gaussian latent space. It is possible a more accurate posterior would be achievable with an adversarial loss function which does not rely on this assumption, but such an approach comes with the additional cost of difficulty in training. In addition, there is the added possibility of biasing the cosmological results due to the adversarial function implicitly learning the cosmology of the training set. For example, the adversarial loss function could implicitly learn to calculate a power-spectra-like function, which would force the generative network to always produce maps with the the same power spectra as the training samples. With an L1 norm we are hopeful that our network would be transferable to other cosmological models, as these models would affect only the dark matter field without appreciably altering the hydrodynamical mapping.

We expect the overall design of the network to be easily extended to other hydrodynamical properties. Simulations suites such as CAMELS (Villaescusa-Navarro et al. 2020), which include parametric feedback models, would allow the creation of an all-purpose mapping tool from dark matter to hydrodynamical models conditioned on underlying physics and redshift. While in NYX the main source of effective stochastic processes are gas shocks, other hydrodynamical simulation tools like AREPO (Springel 2010; Weinberger et al. 2020), used in CAMELS, allow for feedback processes through star formation, active galactic nuclei, supernova, etc. When studying these phenomena, it becomes very important to include a stochastic component to the network, as done in HYPHY. We plan on further exploring these properties in future works.

We thank Mustafa Mustafa, Stephanie Ger, Chirag Modi, Francois Lanusse, Peter Melchior, Uros Seljak, Charles Williams, Stanley Cox, and Earl Stevens for their helpful insights. We dedicate this work to the late Andre Louis Hicks, whose work is a constant inspiration. 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.

Appendix
Light-cone Generation

We construct the inputs to the HYPHY model to have a redshift label for every pixel, a necessary property to allow one model to learn a range of redshifts and maintain a fully convolutional architecture. While this is suboptimal in terms of memory usage in the examples in the main body of the paper, it allows generation of realistic light cones by varying this index over the box. As the convolutional layers are limited in scope, when mapping a given pixel they will only have input from a very limited range of redshifts. We therefore expect to not generate any significant artifacts due to the redshift variation over the test boxes that is not present in the training sample. Our training sample only has fixed redshift boxes at $z = 2.4$, $z = 3.0$, and $z = 4.0$.

In practice the redshift label is most relevant for the temperature of the IGM; for a given fixed dark matter distribution there is little redshift dependence in the baryon density or velocity field. We demonstrate the construction of such a light cone in Figure 12. In the top two panels there is a clear difference in the overall temperature of the IGM with the temperature rising from $z = 4.0$ to $z = 2.4$. The light cone version in the bottom panel, where we feed a varying redshift index, smoothly interpolates between the two. We plan to further apply this light-cone technique for mock generation and forward-model reconstructions in future works.

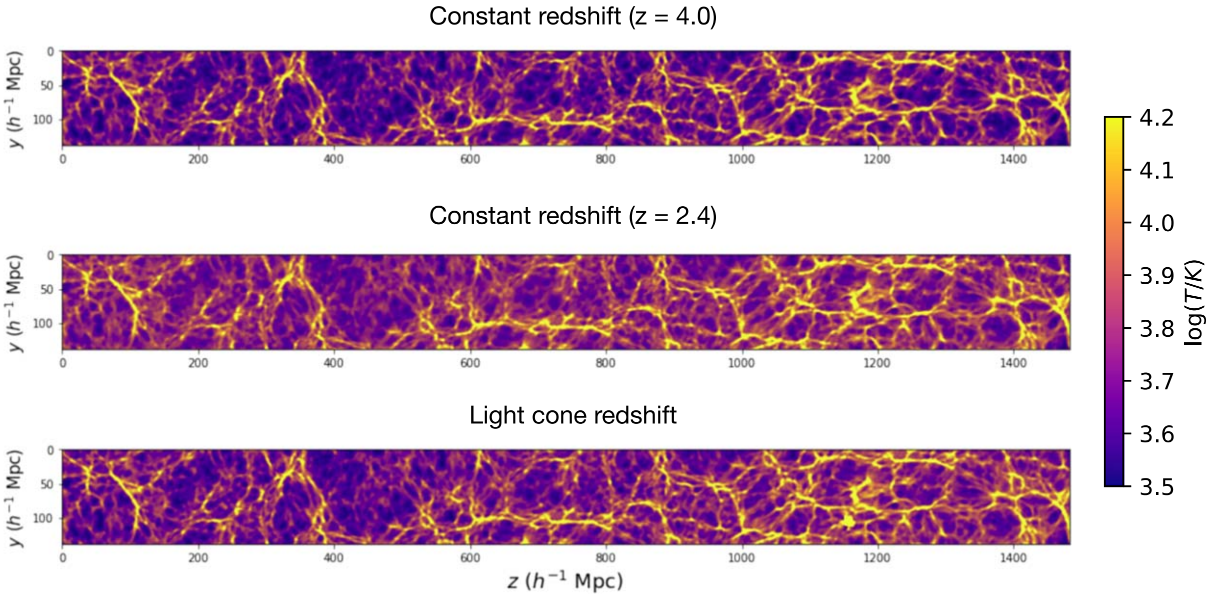

Figure 12. Figure demonstrating the generation of a light cone using HYPHY. In the top two panels, we show the HYPHY-generated temperature corresponding to the same dark matter distribution as though it was at $z = 2.4$ and $z = 4.0$. In the bottom panel we show the same distribution but with a varying redshift across the box, going from $z = 4.0$ (left) to $z = 2.4$ (right).

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. 1992, ApJS, 78, 341

Chabanier, S., Palanque-Delabrouille, N., Yèche, C., et al. 2019, JCAP, 2019, 017

Coc, A., Uzan, J.-P., & Vangioni, E. 2013, arXiv:1307.6955

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

Esser, P., Sutter, E., & Ommer, B. 2018, in Proc. IEEE Conf. Computer Vision and Pattern Recognition (Los Alamitos, CA: IEEE Computer Society Press), 8857

Evrard, A. E. 1990, ApJ, 363, 349

Friesen, B., Almgren, A., Lukić, Z., et al. 2016, ComAC, 3, 4

Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998, ApJ, 503, 518

Galárraga-Espinosa, D., Aghanim, N., Langer, M., & Tanimura, H. 2021, A&A, 649, A117

Gu, J., Wang, Z., Kuen, J., et al. 2018, PatRe, 77, 354

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

Haardt, F., & Madau, P. 2012, ApJ, 746, 125

Harrington, P., Mustafa, M., Dornfest, M., Horowitz, B., & Lukić, Z. 2022, ApJ, 929, 160

Horowitz, B., Lee, K.-G., White, M., Krolewski, A., & Ata, M. 2019, ApJ, 887, 61

Horowitz, B., Zhang, B., Lee, K.-G., & Kooistra, R. 2021, ApJ, 906, 110

Hui, L., & Gnedin, N. Y. 1997, MNRAS, 292, 27

Johnson, J., Alahi, A., & Fei-Fei, L. 2016, in European Conf. Computer Vision (Cham: Springer), 694

Kamyab, S., Sabzi, R., & Azimifar, Z. 2019, in 2019 4th Int. Conf. Pattern Recognition and Image Analysis (IPRIA) (Piscataway, NJ: IEEE), 257

Katz, N., Weinberg, D. H., & Hernquist, L. 1996, ApJS, 105, 19

Khan, S. H., Hayat, M., & Barnes, N. 2018, arXiv:1804.10323

Khoshgoftaar, T. M., Seiffert, C., Van Hulse, J., Napolitano, A., & Folleco, A. 2007, in Sixth Int. Conf. Machine Learning and Applications (ICMLA 2007) (Piscataway, NJ: IEEE), 348

Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980

Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. 2014, in Advances in Neural Information Processing Systems 27, ed. Z. Ghahramani et al. (Red Hook, NY: Curran Associates, Inc.), 3581

Kingma, D. P., & Welling, M. 2013, arXiv:1312.6114

Kiyoiti dos Santos Tanaka, F. H., & Aranha, C. 2019, arXiv:1904.09135

Kullback, S. 1997, Information Theory and Statistics (New York: Dover)

Lanusse, F., Mandelbaum, R., Ravanbakhsh, S., et al. 2021, MNRAS, 504, 5543

Lukić, Z., Stark, C. W., Nugent, P., et al. 2015, MNRAS, 446, 3697

Martizzi, D., Vogelsberger, M., Artale, M. C., et al. 2019, MNRAS, 486, 3766

Modi, C., Feng, Y., & Seljak, U. 2018, JCAP, 10, 028

Palanque-Delabrouille, N., Yèche, C., Schöneberg, N., et al. 2020, JCAP, 2020, 038

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6

Rahmati, A., Pawlik, A. H., Raičević, M., & Schaye, J. 2013, MNRAS, 430, 2427

Rogers, K. K., & Peiris, H. V. 2021, PhRvL, 126, 071302

Ronneberger, O., Fischer, P., & Brox, T. 2015, in Int. Conf. Medical Image Computing and Computer-assisted Intervention (Berlin: Springer), 234

Sohn, K., Lee, H., & Yan, X. 2015, in Advances in Neural Information Processing Systems 28, ed. C. Cortes et al. (Red Hook, NY: Curran Associates, Inc.), 3483

Sorini, D., Oñorbe, J., Lukić, Z., & Hennawi, J. F. 2016, ApJ, 827, 97

Springel, V. 2005, MNRAS, 364, 1105

Springel, V. 2010, MNRAS, 401, 791

Tröster, T., Ferguson, C., Harnois-Déraps, J., & McCarthy, I. G. 2019, MNRAS, 487, L24

Tsang, B. T. H., & Schultz, W. C. 2019, ApJL, 877, L14

Ursino, E., Galeazzi, M., & Roncarelli, M. 2010, ApJ, 721, 46

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2020, ApJ, 915, 71

Wadekar, D., Villaescusa-Navarro, F., Ho, S., & Perreault-Levasseur, L. 2020, ApJ, 916, 42

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

Wang, S., Liu, W., Wu, J., et al. 2016, in 2016 Int. Joint Conf. Neural Networks (IJCNN) (Piscataway, NJ: IEEE), 4368

Weinberger, R., Springel, V., & Pakmor, R. 2020, ApJS, 248, 32

Zamudio-Fernandez, J., Okan, A., Villaescusa-Navarro, F., et al. 2019, arXiv:1904.12846