The Astrophysical Journal, 970:174 (12pp), 2024 August 1 https://doi.org/10.3847/1538-4357/ad5957

Debiasing with Diffusion: Probabilistic Reconstruction of Dark Matter Fields from Galaxies with CAMELS

Victoria Ono[1,2], Core Francisco Park[1,2], Nayantara Mudur[1,2], Yueying Ni[1], Carolina Cuesta-Lazaro[1,3,4], and Francisco Villaescusa-Navarro[5,6]
[1] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA; victoriaono@college.harvard.edu
[2] Department of Physics, Harvard University, 17 Oxford Street, Cambridge, MA 02138, USA
[3] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[6] Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA
Received 2024 March 22; revised 2024 June 12; accepted 2024 June 15; published 2024 July 30

Abstract

Galaxies are biased tracers of the underlying cosmic web, which is dominated by dark matter (DM) components that cannot be directly observed. Galaxy formation simulations can be used to study the relationship between DM density fields and galaxy distributions. However, this relationship can be sensitive to assumptions in cosmology and astrophysical processes embedded in galaxy formation models, which remain uncertain in many aspects. In this work, we develop a diffusion generative model to reconstruct DM fields from galaxies. The diffusion model is trained on the CAMELS simulation suite that contains thousands of state-of-the-art galaxy formation simulations with varying cosmological parameters and subgrid astrophysics. We demonstrate that the diffusion model can predict the unbiased posterior distribution of the underlying DM fields from the given stellar density fields while being able to marginalize over uncertainties in cosmological and astrophysical models. Interestingly, the model generalizes to simulation volumes ≈500 times larger than those it was trained on and across different galaxy formation models. The code for reproducing these results can be found at https://github.com/victoriaono/variational-diffusion-cdm.

Unified Astronomy Thesaurus concepts: Galaxies (573); Cosmology (343); Large-scale structure of the universe (902)

1. Introduction

Within the standard ΛCDM paradigm, hierarchical structure formation arises from the gravitational growth and collapse of the initial density inhomogeneities and forms the large-scale cosmic web of today's universe. In all, 85% of the matter content is composed of dark matter (DM), whose nature remains one of the most enigmatic questions in astrophysics due to the absence of direct observations. Large observational surveys such as the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016), Euclid (Laureijs et al. 2011), Roman (Spergel et al. 2015), and Rubin (LSST Science Collaboration et al. 2009) are devoted to mapping the cosmos by observing millions of galaxies at different wavelengths, which will serve as biased tracers of the underlying DM density fields with the goal of improving our understanding of the nature and constituents of the Universe.

The cosmic web is sensitive to the laws and constituents of the Universe. However, the dominant component of the cosmic web, DM, is not directly observable. Therefore, one needs to infer the distribution of the cosmic web solely based on biased observable tracers. In this paper, we present a probabilistic approach to reconstruct DM density fields from such observations. The reconstructed cosmic web can be used for two purposes: (1) to study field-level cosmology (Cuesta-Lazaro & Mishra-Sharma 2023; Nguyen et al. 2024) and extract information from voids and filaments that may not be detectable from the galaxy distribution alone; (2) to identify regions with low ratios of stellar-to-DM mass, as those are the regions where we expect DM signatures to be larger and can aid in constraining the nature of DM.

On large scales, the clustering of galaxies can be described by a perturbative bias expansion of the DM density (Desjacques et al. 2018), where the complexity of galaxy formation physics is contained in a small set of expansion coefficients referred to as biased parameters. Alternatively, one can describe galaxy biasing in the context of the Effective Field Theory of Large Scale Structures (Senatore 2015), which provides a systematic framework for modeling galaxy clustering based only on symmetries and scale separation. Currently, these models can only accurately reproduce the galaxy power spectrum on scales larger than $10\,h^{-1}$ Mpc (Ivanov 2021).

On smaller scales, perturbation theory breaks down, and the clustering of galaxies is affected by nonlinear structure formation and astrophysical processes such as supernova or active galactic nucleus (AGN) feedback, requiring hydrodynamical simulations for theoretical predictions in this regime. Over the past decade, galaxy formation simulations have made significant progress (see Vogelsberger et al. 2020 for a recent review) to more accurately study the relationship between the observed galaxy distributions and the underlying DM fields.

However, due to limited resolution, cosmological simulations usually adopt coarse-grained subgrid models to effectively describe small-scale astrophysical processes such as star formation, supernova feedback, black hole evolution, and AGN feedback, and there are still large theoretical uncertainties lying

in those subgrid models that lead to different predictions made by different galaxy formation simulations.

To encapsulate those uncertainties, the Cosmology and Astrophysics with Machine Learning Simulations (CAMELS) project (Villaescusa-Navarro et al. 2021; Ni et al. 2023) generated more than 8000 state-of-the-art galaxy formation simulations that widely explore the variations in cosmological and astrophysical parameters within different galaxy formation models to provide a broad training set and testing ground for machine-learning (ML) algorithms designed for cosmological studies.

ML is a powerful tool to learn complex high-dimensional distributions, such as the mapping between observable fields (e.g., galaxies, neutral hydrogen gas) and the underlying cosmic web from simulations. For example, previous work has tried to use convolutional neural network (NN) models to infer the DM field from the galaxy density field (Hong et al. 2021) and from 21 cm maps (Villanueva-Domingo & Villaescusa-Navarro 2021). However, these models are deterministic and cannot generate the posterior distribution of different samples of the cosmic web given the observable fields. It is important to develop probabilistic generative models, such as those of Mudur & Finkbeiner (2022) and Park et al. (2023), which can encapsulate our uncertainties of how galaxies connect to the underlying DM distribution. These models enable us to address questions such as the likelihood of a certain DM halo mass being associated with a given galaxy distribution.

In this work, we develop a diffusion generative model trained on the CAMELS simulation suites to reconstruct the underlying DM fields from stellar density fields. The primary goal of the diffusion model is to capture the relationship between the stellar fields and DM fields and predict the unbiased posterior distribution of the DM fields conditioned on the given stellar field, $p(x_{\rm DM}|x_{\rm stars})$. By training on the CAMELS suite, the diffusion model can account for the uncertainties inherent in the astrophysical processes assumed by galaxy formation models, as well as the specific cosmological parameters utilized in the simulations. We also train the diffusion model using various training sets from CAMELS and evaluate its robustness across different galaxy formation models. Additionally, we apply the trained diffusion model to large simulations of IllustrisTNG-300 to showcase its extrapolation performance on out-of-distribution data.

The paper is structured as follows. Section 2 presents the diffusion model along with the data set from CAMELS used for training and testing purposes. Section 3 conducts a series of validations to assess the performance and robustness of the diffusion model. Finally, Section 4 provides a summary of the paper.

2. Methodology

In this section, we describe the data set used, the model architecture, and training methods.

2.1. Data Set

In this work, we use 2D maps from the CAMELS Multifield Data Set (Villaescusa-Navarro et al. 2022) to model the connection between stellar mass and DM densities produced by three different suites of hydrodynamical simulations: ASTRID, IllustrisTNG, and SIMBA. Table 1 provides a summary of the simulations in each suite.

CAMELS simulations have a box volume of $(25h^{-1}{\rm Mpc})^3$. Each simulation evolves the universe from $z=127$ to today, following the evolution of 256^3 DM particles of mass $6.49\times10^7(\Omega_m-\Omega_b)/0.251\,h^{-1}\,M_\odot$ and 256^3 gas resolution elements with an initial mass of $1.27\times10^7\,h^{-1}\,M_\odot$. From each simulation at $z=0$, the CAMELS Multifield Data Set produces 15 paired maps representing the stellar and DM surface density in a region with dimensions $25\times25\times5\,(h^{-1}{\rm Mpc})^3$ that is projected along the third axis. We keep the image size to the original 256×256 pixels, corresponding to $25\,h^{-1}$ Mpc on both sides. The resolution of our maps is therefore of $\approx0.1\,h^{-1}$ Mpc.

We use the Latin hypercube (LH) set of each of the three simulation suites to train our models. Each LH sets contains 1000 independent simulations spanning a wide range of cosmological and astrophysical parameters, reflecting the uncertainties of cosmology and the complex astrophysical processes taking place in our current understanding of galaxy formation. We augment the training set with random flips and permutations of the input and output images.

The parameters varied in the LH set for each simulation suite are $\Omega_{\rm m}$, σ_8 (cosmological) and $A_{\rm SN1}$, $A_{\rm AGN1}$, $A_{\rm SN2}$, and $A_{\rm AGN2}$ (astrophysical), and their ranges are as follows: $0.1\leqslant\Omega_m\leqslant0.5$, $0.6\leqslant\sigma_8\leqslant1.0$, $0.25\leqslant(A_{\rm SN1},A_{\rm AGN1})\leqslant4.00$, and $0.5\leqslant(A_{\rm SN2},A_{\rm AGN2})\leqslant2.0$. $A_{\rm SN}$ and $A_{\rm AGN}$ control the strength of supernova wind and AGN feedback, respectively, with their exact physical meaning differing across different galaxy formation models. In particular, $A_{\rm SN1}$ stands for the energy (in ASTRID and IllustrisTNG) or mass loading factor (in SIMBA) of galactic winds per unit star formation rate; $A_{\rm SN2}$ controls the speed of the galactic wind; $A_{\rm AGN1}$ variates the energy (in ASTRID and IllustrisTNG) or momentum (in SIMBA) of the AGN jet mode feedback; $A_{\rm AGN2}$ controls the energy of the AGN thermal mode feedback in ASTRID, the jet speed and burstiness in IllustrisTNG, and the jet speed in SIMBA. We refer to Ni et al. (2023) for more detailed descriptions.

We use the different simulation suites to assess the generalization capabilities of the trained diffusion models. In particular, the 1P set varies one parameter (from the fiducial value in the CV set) at a time for each simulation. Testing on this data set would clearly indicate how capable the model is at marginalizing over each astrophysical parameter. On the other hand, the CV set contains 27 simulations with the same fiducial values of cosmological and astrophysical parameters but also varied random seeds of the initial conditions that are designed to quantify the level of cosmic variance on different cosmological and astrophysical probes. The fiducial values are $\Omega_m=0.3$, $\sigma_8=0.8$, $A_{\rm SN1}=A_{\rm AGN1}=A_{\rm SN2}=A_{\rm AGN2}=1$.

2.2. Diffusion Model

Figure 1 provides a schematic overview of the diffusion model used in this study. We use the variational diffusion model developed by Kingma et al. (2021) with a denoising architecture similar to the U-Net (Ronneberger et al. 2015) to model the posterior distribution of DM density fields from the given stellar fields. The conditional diffusion model predicts the conditional probability of the DM field $x_{\rm DM}$ from the stellar field $x_{\rm stars}$:

$$p(x_{\rm DM}|x_{\rm stars})=\frac{p(x_{\rm DM})p(x_{\rm stars}|x_{\rm DM})}{p(x_{\rm stars})}.$$

Our diffusion model generates a target DM density field in $T=250$ refinement steps. In the forward diffusion process, we

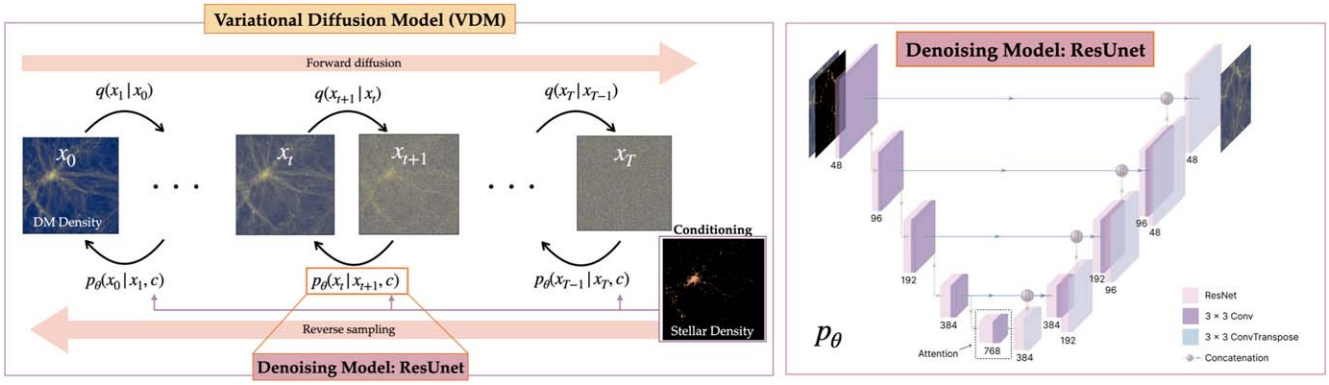

Figure 1. Schematic overview of the conditional diffusion model used to model the posterior distribution of DM density fields from the given the stellar density fields. The left panel illustrates the diffusion process, and the right panel shows the details of the convolutional NN-based denoising model.

Table 1
Summary of the Simulation Sets, Which Share the Same Design between the ASTRID, IllustrisTNG, and SIMBA Suites

Simulation Suites	Set	Number of Simulations	Ω_m	σ_8	$A_{\rm SN1}$	$A_{\rm SN2}$	$A_{\rm AGN1}$	$A_{\rm AGN2}$
ASTRID / IllustrisTNG / SIMBA	LH	1000	0.1–0.5	0.6–1.	0.25–4.	0.5–2.	0.25-4.	0.5–2.
...	1P	61	0.1–0.5	0.6–1.	0.25–4.	0.5–2.	0.25-4.	0.5–2.
...	CV	27	0.3	0.8	1	1	1	1

Note. $A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, and $A_{\rm AGN2}$ represent the value of subgrid physics parameters controlling stellar and AGN feedback. The LH set varies parameters in a Latin hypercube, whereas the 1P set has variations of each parameter at a time, with all others fixed to their fiducial values. The CV set has all parameters fixed to the fiducial values but varies the random seed of the initial conditions.

progressively add noise to an image by sampling from $q(x^t_{\rm DM}|x^0_{\rm DM}) = \mathcal{N}(\alpha_t x^0_{\rm DM}, \sigma_t^2 \boldsymbol{I})$, where α_t and σ_t^2 are functions of γ_t, the variance schedule, which we assume to be a linear function of time and whose free parameters are learned during training. Noise is added to the sample according to this schedule and in a variance-preserving way, i.e., $\alpha_t^2 = {\rm sigmoid}(-\gamma(t))$.

In the reverse diffusion process, we wish to sample from $q(x^{t-1}_{\rm DM}|x^t_{\rm DM})$ to denoise the image until we reproduce the original image, but this is intractable as it requires knowing the entire distribution of all possible images. Thus we use an NN to estimate the conditional probability $p_\theta(x^{t-1}_{\rm DM}|x^t_{\rm DM}, x_{\rm stars})$, where $x_{\rm stars}$ is the corresponding star image to $x_{\rm DM}$ called a conditioning. It begins with a random Gaussian noise field $x^T_{\rm DM} \sim \mathcal{N}(0, 1)$ and iteratively denoises it according to this learned probability to ultimately generate a sample $x^0_{\rm DM} \sim p(x_{\rm DM}|x_{\rm stars})$. During training, the denoising NN takes as input $\{x^t_{\rm DM}, x_{\rm stars}, t\}$ and estimates the noise that was added to the image at that time step.

The loss function we optimize is the variational lower bound of the marginal likelihood,

$$L_{\rm VLB} = \underbrace{D_{\rm KL}(q(x^1_{\rm DM}|x^T_{\rm DM})\|p_\theta(x^1_{\rm DM}))}_{\text{Prior loss}} + \underbrace{\mathbb{E}_{q(x^0_{\rm DM}|x^T_{\rm DM})}[-\log p_\theta(x^T_{\rm DM}|x^0_{\rm DM}, x_{\rm stars})]}_{\text{Reconstruction loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}(0,1)}[{\rm SNR}'(t)\|x^T_{\rm DM} - \hat{x}^t_{\rm DM}\|_2^2]}_{\text{Diffusion loss}},$$

where ${\rm SNR}'(t)$ is the derivative of $\exp(-\gamma(t))$, the signal-to-noise ratio as a function of time that we assume to be decreasing in time, and $\hat{x}^t_{\rm DM}$ is the predicted output from the denoising model at time step t. The objective thus is to minimize a bound to the posterior $p_\theta(x_{\rm DM}|x_{\rm stars})$.

A detailed ablation study of this model is presented in Appendix B.

2.3. Training

For our denoising model, we use a hierarchical U-Net (Ronneberger et al. 2015)-like architecture with four blocks of double convolution followed by strided downsampling layers. We employ group normalization (Wu & He 2018) and residual connections (He et al. 2015) in each block and use the AdamW optimizer (Loshchilov & Hutter 2017) with a learning rate of 1×10^{-4} and the CosineAnnealingWarmRestarts learning rate scheduler (Loshchilov & Hutter 2016). We also initialize the learned linear noise schedule with $\gamma(t) = 26.6\,t - 13.3$. We train the model using the PyTorch Lightning framework (Falcon et al. 2020) with a batch size of 12 for 60000 gradient steps, using the LH set as our training data.

We select the model with the lowest mean-squared error (MSE) in the validation set, calculated as

$${\rm MSE} = \mathbb{E}_{x^{\rm Sample}_{\rm DM}\sim p(x_{\rm DM}|x_{\rm stars})}(x^{\rm Sample}_{\rm DM} - x^{\rm True}_{\rm DM})^2.$$

3. Results

In this section, we conduct an exhaustive set of tests to demonstrate the trained model's performance and robustness:

1. In Section 3.1, we showcase the model's performance when the cosmological and astrophysical parameters are set to their fiducial values, using the CV set for testing the model.
2. In Section 3.2, we demonstrate that the model generalizes to different cosmological and astrophysical parameters by varying one of them at a time, using the 1P set.

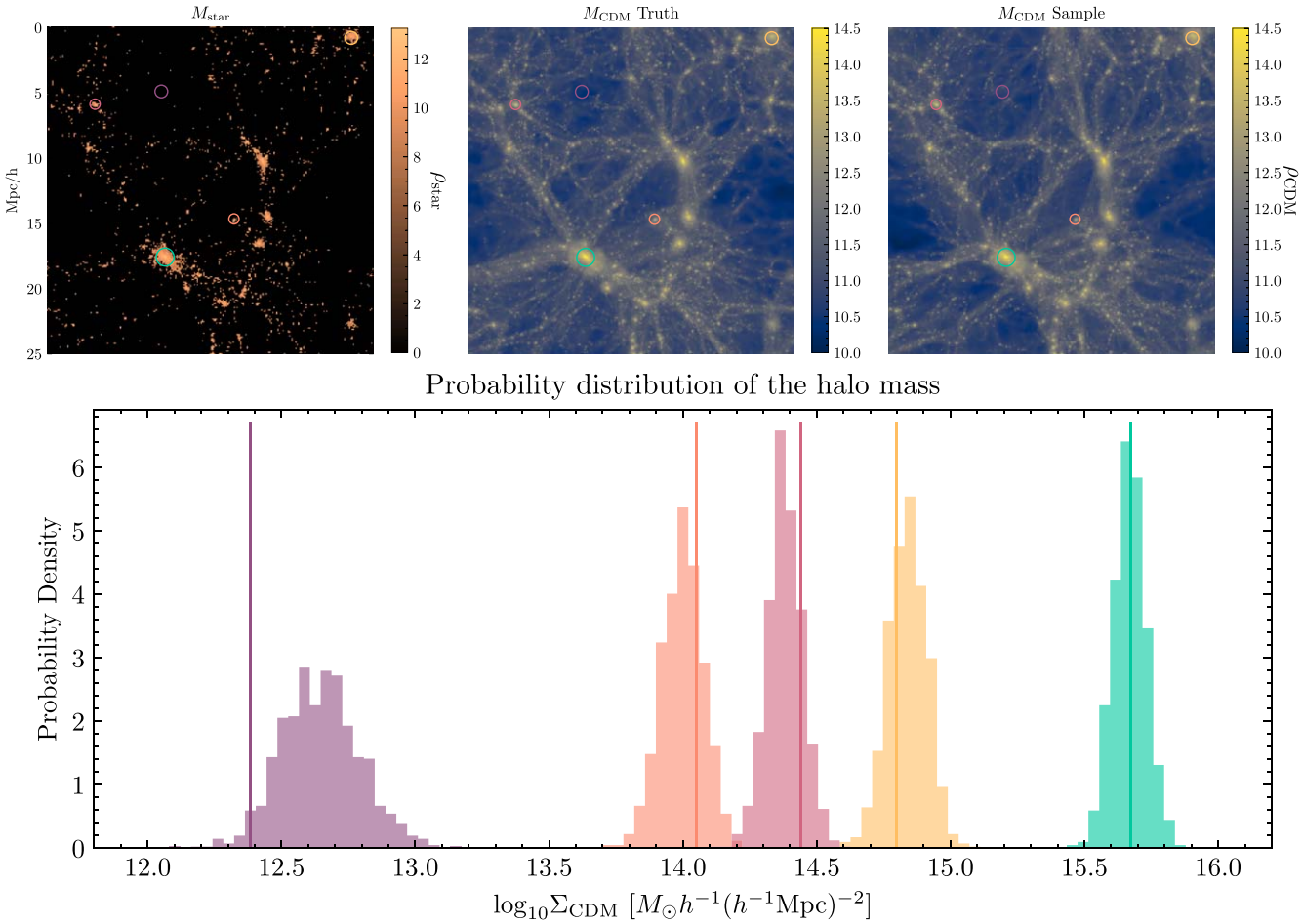

Figure 2. Top row: input stellar field, corresponding true DM density field, and a sample DM density field from the ASTRID-trained model, with circles drawn for selected regions ranging from a void to a massive star cluster. Bottom row: probability distribution of the DM mass (defined to be the sum of the pixels in the region) from 1000 samples corresponding to each selected region, with the solid line representing the true mass. The model is able to predict closer to the true mean as the total mass in the region increases, and its decrease in variance demonstrates the model's increase in confidence.

3. In Section 3.3, we show that the model generalizes across galaxy formation simulations after training the diffusion model on a single simulation suite and testing on the others.
4. In Section 3.4, we test the model's ability to recover the large-scale DM distribution by deploying a model trained on small volumes to a simulation with an ≈500 times larger volume.

3.1. Predictions for the CV Set

In this section, we test the diffusion model trained on maps of the LH set of the ASTRID simulation suite on maps from ASTRID's CV set, whose cosmological and astrophysical parameter values are fixed to the fiducial ones.

The top panel of Figure 2 illustrates one posterior sample of the DM density field in comparison to the true DM density field. The model can qualitatively reproduce the expected features of the DM cosmic web of nodes, filaments, and voids.

The bottom panel in Figure 2 demonstrates how the posterior samples can be used to estimate posteriors of quantities of interests, in particular the mass of selected regions in the density field. Here, the mass is calculated as the sum of pixel values in the selected circular region. The regions for which posteriors are shown are highlighted in the maps above, ranging from a void to a massive star cluster. The radius of each region is determined such that it encompasses sufficient mass to illustrate the probability distribution as applying a halo finder would require 3D density fields, while we use projected 2D fields. As expected, the posteriors are broader in regions with low density of stars.

We also train the same diffusion model on the CV set instead of the LH set to test how much uncertainty is introduced in the posterior samples by implicitly marginalizing over the varying astrophysical and cosmological parameters in the LH set. We do not see a significant reduction in variance in the model trained on the CV set given the large range of variations in the cosmological and astrophysical priors used to generate the LH set simulations. We hypothesize this small difference is due to the ability of the LH-trained model to infer the correct cosmological parameters from the stellar mass maps to debias the fields, as shown in Section 3.2. This reduction in variance would also come at the cost of poorer generalization capabilities. We therefore train all our models on the LH set.

Figure 3 shows a quantitative comparison between the summary statistics of the true DM fields and those of the generated ones from the corresponding stellar fields. The results are shown for one sample of the CV set. From left to right, we show the statistics of density histograms, the 2D power spectra, and the 2D cross power spectra, with shaded

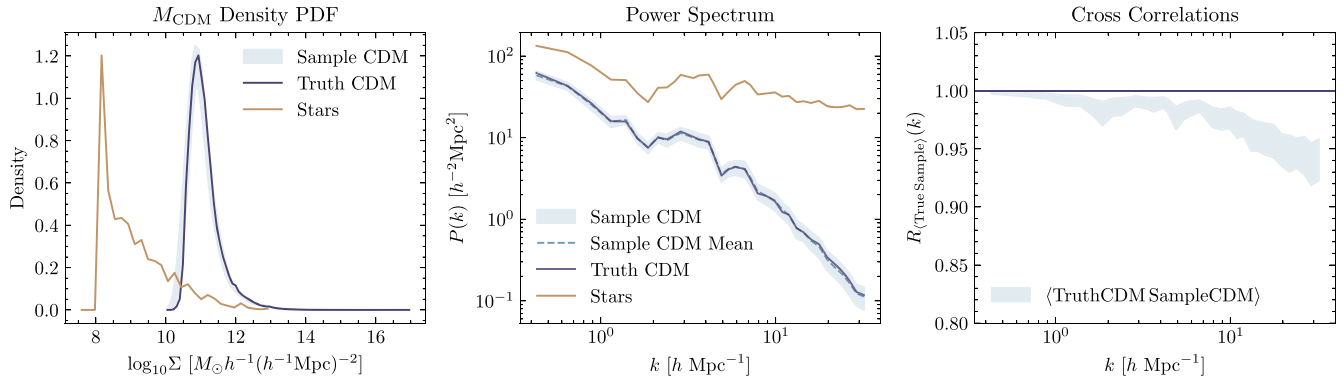

Figure 3. Summary statistics of the single stellar field and its corresponding DM density field from the CV set. Left panel: density histogram of star (only nonzero stellar densities in the 2D maps are shown in the copper line), true DM (solid blue line), and DM inferred by the diffusion model (light blue). The shaded region shows the 10th–90th percentiles of 100 samples from the posterior distribution of the DM field, and the dashed line shows its mean. Middle panel: power spectra for star, true DM, and sampled DM fields. Right panel: cross-correlations between true and sampled DM fields. All panels show good agreement between the summary statistics of the true and sampled DM fields, demonstrating the model's ability to well reproduce the statistical properties of the cosmic web.

regions obtained from the posterior distribution of 100 samples of the generated DM map, quantifying the 10th to 90th percentile uncertainties of the model predictions.

The left panel of Figure 3 shows good agreement between the 1D histogram of the true DM field and the sampled DM fields.

The middle panel of Figure 3 compares the power spectra of the true DM map and generated samples. We see that the diffusion model can reproduce the expected dependence of the power spectrum with scale and that the true power falls within the uncertainty of the samples. The relatively scale-independent variance of the posterior samples arises from degeneracies in the DM density field that can lead to similar stellar distributions after marginalizing over cosmological and astrophysical parameters.

Finally, the right panel of Figure 3 compares the cross-correlation coefficient between the true and sampled DM fields. Cross-correlation coefficients are calculated by taking the cross-power spectra between the sampled DM and true DM field and dividing by their autopower spectra:

$$R_{\langle \rm True\ Sample\rangle}(k) = \frac{P_{\langle \rm True\ Sample\rangle}(k)}{\sqrt{P_{\rm True}(k)}\sqrt{P_{\rm Sample}(k)}},$$

where $R(k)$ measures the correlation between the phases of the modes of the two fields. In an ideal scenario where the true DM and sampled DM fields are perfectly correlated, the cross-correlation between true DM and sample DM would be 1 across all the scales. However, perfect reconstruction of the DM field is not possible due to limitations in the information contained within the stellar mass maps, such as the discrete nature of the tracer (shot noise). We find that the cross-correlations between the sampled and true DM fields are always higher than 0.9, demonstrating that the model is able to well reproduce the statistical properties of the cosmic web. This value serves as a lower bound on the information content of the stellar mass fields about the DM density field. Different architectures, data sets, or optimization methods may be able to reconstruct the DM cosmic web more precisely, potentially improving upon the cross-correlation values we have obtained.

In Appendix C, we show how the posterior variance is lowest (highest) at pixels with high (low) nonzero stellar mass.

3.2. Varying Cosmological and Astrophysical Parameters

Figure 4 shows a quantitative comparison between the generated DM density fields and the true ones based on the simulations in the 1P set that separately vary the parameters Ω_m, $A_{\rm SN1}$, and $A_{\rm AGN1}$. For each selected parameter, we randomly choose a single stellar map from that simulation as the input to generate 100 DM samples.

As shown in the first column of the upper panel, the DM distribution is more sensitive to the cosmological parameter Ω_m. We show that in both low- and high-Ω_m scenarios, the diffusion model can effectively capture the dependence on Ω_m and reproduce the trends in the DM density distribution accordingly.

In the lower panel of Figure 4, we show the ratio of the predicted power spectra to the true one. We find that the model tends to overpredict the power across all scales for large variations in Ω_m for this particular sample, although this is not a general trend, as can be seen in Figure 3.

For varying $A_{\rm SN1}$ and $A_{\rm AGN1}$ scenarios, the stellar density maps will look very different on small scales due to the varying strength in supernova and AGN feedback, while the underlying true DM density field will be largely unaffected. We see that the model is able to reproduce this behavior and generalizes well for both low and high parameter values, exhibiting its potential to marginalize over baryonic effects. We do, however, see more uncertainty at smaller scales, consistent with our findings from other summary statistics.

The consistency of the density probability distribution functions (PDFs) and power spectra with varied cosmological and astrophysical parameters demonstrates that the trained diffusion model can well capture the clustering properties of the underlying DM fields based on the stellar fields while marginalizing over the cosmological and astrophysical parameters.

Finally, we assess the cosmological information contained in the reconstructed DM density fields. In particular, we use the parameter inference networks presented in Villaescusa-Navarro et al. (2021) to predict the cosmological parameters Ω_m and σ_8 for the reconstructed DM density fields. The network is trained to return a mean prediction as well as a standard deviation that indicates the network's uncertainty on its prediction.

For this test, we select two input stellar fields for each parameter in the 1P set with Ω_m and σ_8 varying and generate 10

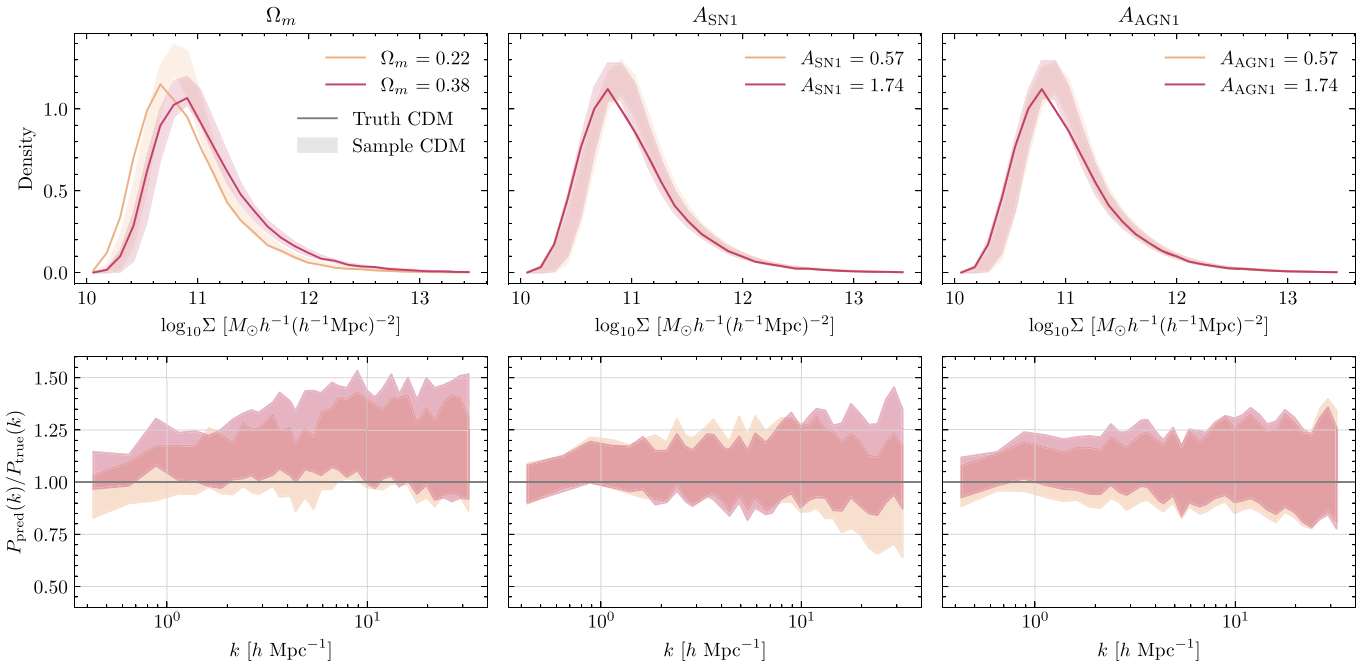

Figure 4. Summary statistics of the density fields when varying Ω_m, $A_{\rm SN1}$, and $A_{\rm AGN1}$ while fixing all other parameters to their fiducial values. For each parameter variation, we use one single stellar map as the input and generate 100 DM samples from that map. Top row: comparison of the density field PDFs. Bottom row: power spectrum ratio between the sampled fields and the true DM field. The solid lines in the top panel give the true DM distributions from simulations, while the shaded regions correspond to 10th–90th percentiles based on 100 generated DM samples from the single input stellar field. Note that the DM projected density distribution does not vary noticeably with the astrophysical parameters $A_{\rm SN1}$ and $A_{\rm AGN1}$, and the two solid lines overlap. The model is able to capture the overall trends well for all three parameters, though its uncertainty increases at smaller scales.

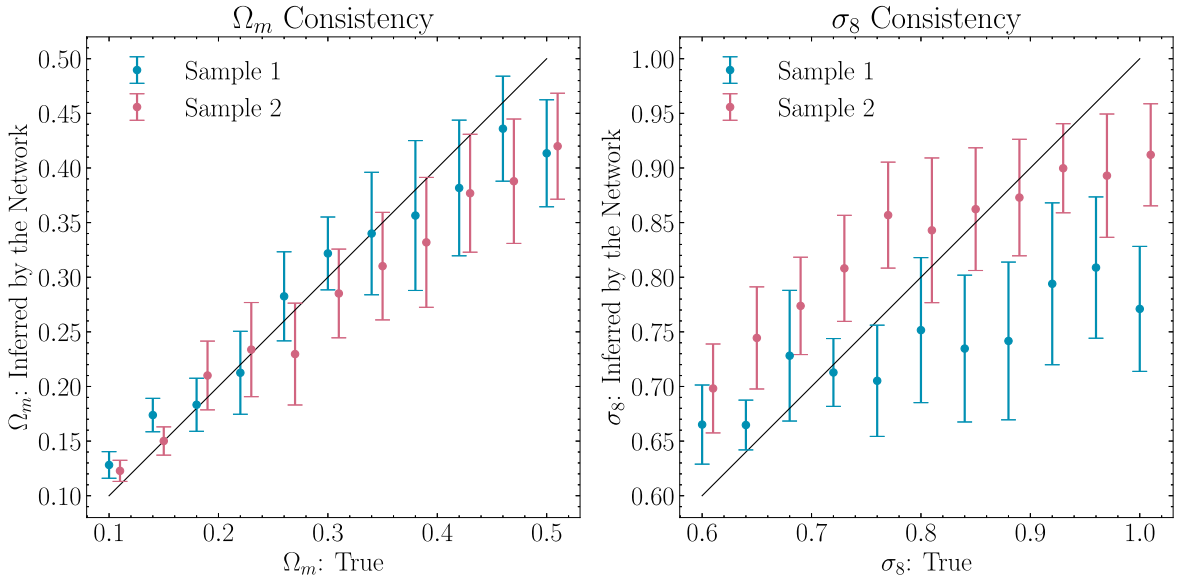

Figure 5. Predictions of the parameter inference network in Villaescusa-Navarro et al. (2021) for 10 samples from the diffusion model $p_\theta(x_{\rm DM}|x_{\rm stars})$ for two independent stellar fields, denoted by the blue and red colors, for each of the 22 1P parameters with varying Ω_m and σ_8. We add a slight offset on the x-axis to help distinguish the two input fields. Each point and its error bars denote the mean and the standard deviation of the NN's parameter predictions over all 10 posterior samples of the input field corresponding to the true parameter value on the x-axis. The values of Ω_m inferred by the NN have a strong correlation with the true parameters, despite the fact that the diffusion model is not conditioned on cosmology.

samples for each field from $p_\theta(x_{\rm DM}|x_{\rm stars})$. We then pass the generated DM density fields through the parameter inference network and examine the consistency between the parameters inferred by the NN and the true cosmological parameters in Figure 5.

The network-inferred values of Ω_m are close to the truth and have an average error of 14.6% while struggling with the σ_8 prediction. Since the parameter inference networks have been trained on a subset of IllustrisTNG DM LH fields while we were testing our model on the ASTRID 1P set, we have found a slight offset in the inferred value of σ_8 when testing the networks on the true ASTRID 1P fields (see Figure 12 in the Appendix D).

In Figure 2 of Villaescusa-Navarro et al. (2021), the parameter inference network trained on input stellar fields was able to predict Ω_m with an average error of 19.8%, in line with our findings.

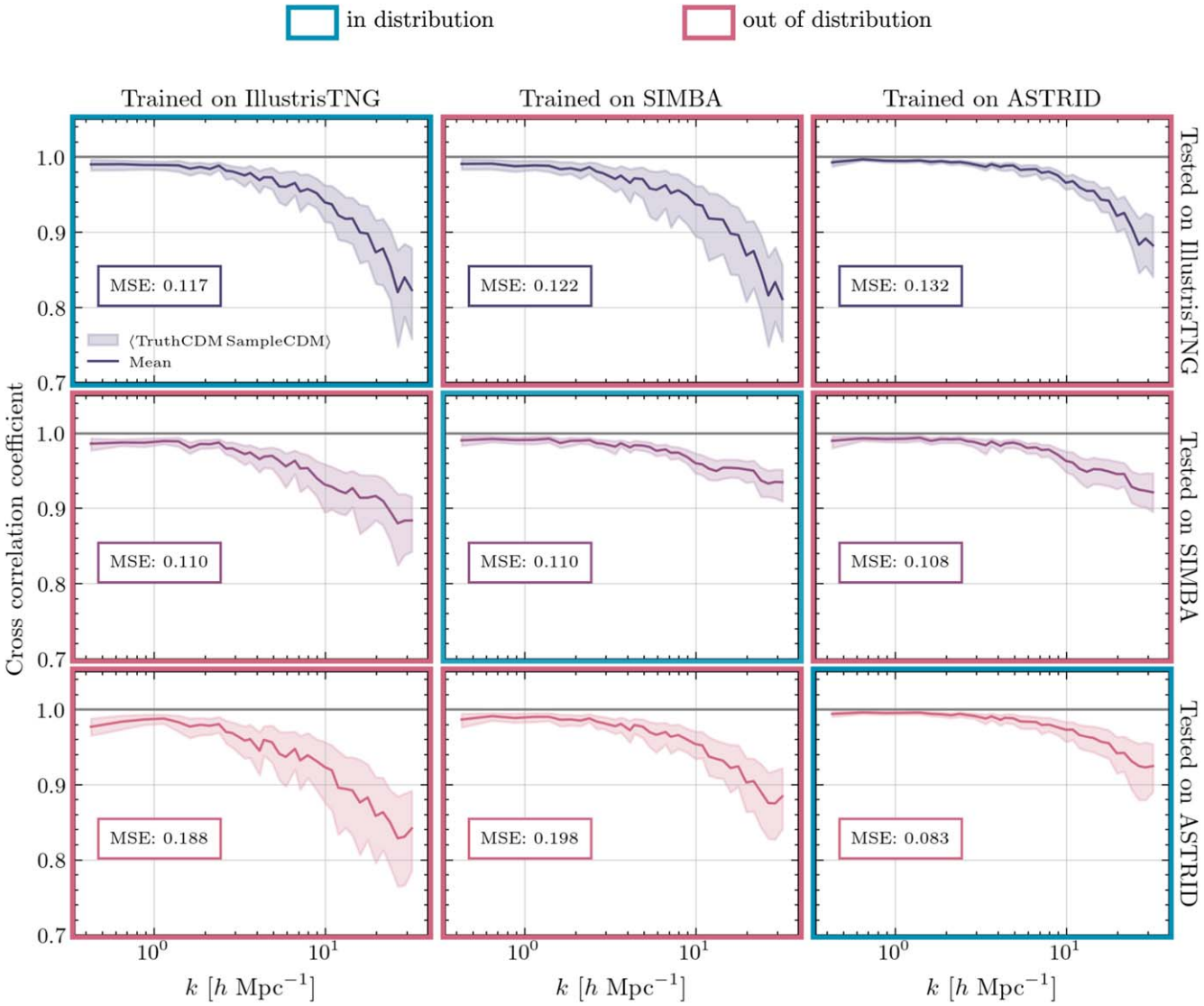

Figure 6. Cross-correlations between true and 100 sampled DM fields given the same conditioning across three models trained and tested on each data set together with the average MSE. The shaded region in each panel represents the 10–90th percentiles of the cross-correlation coefficients for all samples, and the solid line represents their mean coefficient. We observe that the ASTRID-trained model is able to reproduce this quantity the best among all suites, while the IllustrisTNG-trained model produces the lowest coefficients and largest variances. Overall, all different combinations produce cross-correlation coefficients higher than 0.8, demonstrating that each model can generalize well across galaxy formation models.

3.3. Generalizing beyond the Training Set

To test the model's ability to generalize over different galaxy formation simulations, we train three independent models using maps from the IllustrisTNG, SIMBA, and ASTRID LH sets and test each model using maps of the CV set from each of the simulation suites. In Appendix A (Figure 9), we show the generated samples from the same index of each CV set, which corresponds to having the same initial seed and parameter values.[7]

We also quantitatively compare the cross-correlation coefficients between the true DM fields and those of the generated ones for each of these models in Figure 6. As a further comparison metric, we use 100 DM samples generated from the diffusion model and calculate the averaged MSE of the true DM field and the generated samples.

As shown in Figure 6, all three diffusion models perform well when tested on simulations conducted by the same galaxy formation model (the blue boxes marked as in distribution), and they also demonstrate robust generalization when tested on simulations carried out by other galaxy formation models (the red boxes marked as out of distribution). The generated DM fields consistently reproduce the pattern of the true cosmic web across all the scales, maintaining cross-correlation coefficients consistently higher than 0.8 for any pair.

The effective reproduction of DM density fields beyond each model's training set is facilitated by the large overlap of data distribution between the simulation suites, as illustrated in Figure 8 in Appendix A. which shows the ratio of stellar mass to DM density power spectra for each LH set. Despite the distinct galaxy formation models employed by the three simulation suites, resulting in different mappings between the star and DM fields, the overall similarities suggest why each model can effectively generalize to every other galaxy formation model.

[7] Note that the definitions of the astrophysical parameters are different for each simulation suite.

Comparing the models trained on the three different simulation suites, the model trained on the ASTRID suite performs the best with the minimum MSE, while the model trained on IllustrisTNG yields the largest averaged MSE. We speculate that the good performance of the ASTRID-trained model is due to the higher mass resolution of the ASTRID simulation for star particles. In the galaxy formation model employed by ASTRID, each star-forming gas particle can be split and spawn four star particles, as compared to a single star particle in IllustrisTNG and SIMBA. Consequently, the stellar density maps within the ASTRID suite exhibit a less sparse distribution compared to the other two simulation suites, as depicted in the first column of Figure 9 in Appendix A. The effective increased spatial resolution in ASTRID results in stellar density fields that have tighter correlations with the underlying DM fields, explaining the better performance of the ASTRID-trained model.

On the other hand, the relatively poorer performance of the IllustrisTNG-trained model is likely attributed to the fact that IllustrisTNG has suppressed the low-mass galaxy population compared to SIMBA and ASTRID (see, for example, de Santi et al. 2023; Ni et al. 2023), resulting in the most sparse stellar maps and making the training more challenging.

3.4. Recovering the Large-scale Structure in IllustrisTNG300

A downside to training our models on the CAMELS simulations is their limited volume, which in particular will be largely affected by supersample covariance (Li et al. 2014). In this section, we assess the model's capability to generalize to larger volumes than those in the training set by estimating the DM density in the IllustrisTNG300 simulation (Pillepich et al. 2017; Springel et al. 2018; Nelson et al. 2021) that has a volume over 500 times larger than that of the CAMELS simulations used to train the model.

The diffusion model is trained based on the CAMELS data set with an image size of 256^2. However, due to the convolutional nature of the diffusion model applied in this work, the trained model can be applied to fields of different sizes as long as the input fields have the same spatial resolution as that of the training sets. We preprocess the stellar and DM density fields of IllustrisTNG300 by maintaining the same spatial resolution as the CAMELS Multifield Data Set, resulting in projected surface density fields of stars and DM with a size of 2100^2. We then directly apply the trained diffusion model from the CAMELS data set (keeping the same parameters for the convolutional kernels and biases) to the 2D slabs of the IllustrisTNG300 stellar fields. This generates the corresponding DM density fields.

In the upper panel of Figure 7, we show the stellar field of a 2D slab of IllustrisTNG300 together with a sample of the DM density field of the diffusion model trained on the small volume CAMELS-ASTRID suite and the true underlying DM density field from the simulation. The model can surprisingly produce the correct large-scale filaments and voids. Note that we are both extrapolating the trained model to larger volumes and over simulation suites (the model is trained on ASTRID and tested on IllustrisTNG).

In the intermediate panel, we show two highlighted CAMELS-sized regions: a cluster-like region (A) and a void-like region (B). Interestingly, although the void-like region is as large as a CAMELS box, the model still produces a realistic DM density field.

Finally, the bottom panel shows the predicted and true DM density in each pixel, the density PDF, and the power spectra of the samples and the truth, from left to right.

On the left, we show that although the amount of scatter in the predicted versus true relation is considerable, we do not see any bias appearing at high values. This is surprising given that the small volume of the CAMELS simulation limits the appearance of such high-density peaks in the training set.

Moreover, on the bottom-right corner, the power spectra of the DM samples agree very well with those of the true IllustrisTNG300 DM density field across all scales, even recovering large scales that extend beyond the fundamental k-modes corresponding to the small volume training set. For comparison, we also show the star and DM power spectra from the IllustrisTNG CV set, which exhibit a lower amplitude due to the limited large-scale modes in the small-box simulations.

The good agreement of the large volume power spectra demonstrates that the diffusion model (trained on the ASTRID suite) can generalize to (1) different galaxy formation models and (2) scales larger than those contained in the training set.

4. Summary and Discussion

In this paper, we have presented a diffusion generative model that can sample the posterior distribution of DM density fields conditioned on the stellar density field. We demonstrate through a diverse set of validation metrics that the generated DM fields are in good statistical agreement with the true DM fields from the simulations.

Moreover, when trained on the LH set of the CAMELS simulation suites, the diffusion model is able to marginalize over the cosmological and astrophysical uncertainties within a given galaxy formation simulation. Interestingly, the diffusion model exhibits generalization capabilities and can accurately recover DM density fields from simulations with alternative galaxy formation models.

Compared to the previous work of Hong et al. (2021), which applies a deterministic convolutional NN model to learn the mapping between galaxies and DM fields based on the IllustrisTNG simulation, our approach is probabilistic in nature and therefore can capture the inherent uncertainty of the galaxy–DM mapping due to sparsity in the stellar maps, which leads to degeneracies in consistent DM distributions and theoretical uncertainties. This in particular allows us to recover posterior samples with consistent small-scale clustering, whereas deterministic models recover a blurry image smoothed on small scales by training the model to reproduce only the posterior mean.

Notably, the mapping between galaxies and DM fields can also arise from uncertainties in the cosmological parameters and more importantly in subgrid physical models used in galaxy formation simulations. Therefore, we train the diffusion model based on the LH set of CAMELS simulations; this data set features a wide variation in cosmological and astrophysical parameters, allowing the diffusion model to learn how to effectively marginalize over these uncertainties.

We used the 2D projected stellar density field and DM density field from the CAMELS Multifield Data Set as proxies for the galaxy fields and the underlying cosmic web. This approximation is, however, rather simplistic and only serves as a proof-of-concept training set to assess the performance of the diffusion model. To apply the diffusion model to observations of galaxy surveys, future efforts need to focus on making

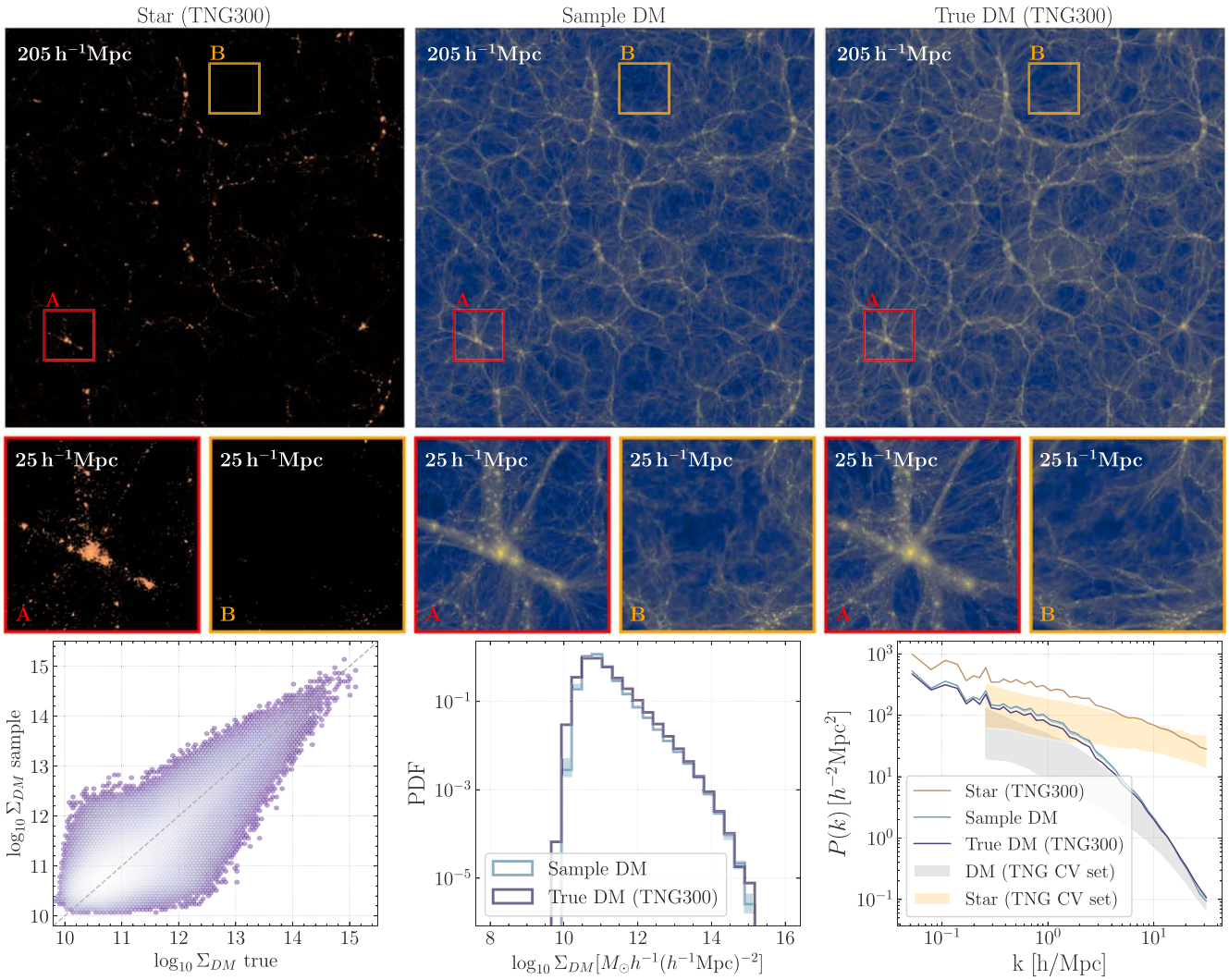

Figure 7. We train the diffusion model on the LH set of the ASTRID suite and apply it to a much larger volume stellar field from the IllustrisTNG300 simulation. Top row: full 205 Mpc h^{-1} stellar density field, the generated DM sample (from our model trained in 25 Mpc h^{-1} CAMELS maps), and the corresponding true DM density field from IllustrisTNG300. The red and orange boxes highlight 25 Mpc h^{-1} subregions of a cluster-like region (A) and a void-like region (B). Bottom row: statistical validations for the 205 Mpc h^{-1} fields illustrating the pixel-level $\Sigma_{\rm DM}$ for both the true and generated DM samples on the left, the 1D histogram of the DM sample and true $\Sigma_{\rm DM}$ in the middle, and the 2D power spectra of the DM and stellar density fields on the right. For comparison, in the rightmost panel we also show the star and DM power spectra from the CAMELS IllustrisTNG CV set, which has a lower amplitude due to missing large-scale modes.

ensembles of simulations with realistic mock synthetic observations of galaxies from simulations, as done in Hahn et al. (2023).

In the future, we plan to develop a training set targeted at reconstructing the DM cosmic web from 3D galaxy point cloud observations of the DESI survey. In particular, 3D galaxy point clouds are sparser than the stellar mass maps used in this work, and processing them could potentially require a prohibitive amount of GPU memory. We plan to assess the feasibility of either doing diffusion in a compressed latent space (Rombach et al. 2021) or developing a diffusion model that can be conditioned on the sparse galaxy 3D point cloud by using either graph NNs or transformers as conditioning models (Cuesta-Lazaro & Mishra-Sharma 2023).

Finally, we will include observational effects, such as selection biases, redshift-space distortions, and fiber collisions, in order to train a generative model that can be effectively applied to real galaxy surveys and unravel the cosmic web of our Universe.

Acknowledgments

We thank Daniel Eisenstein and Lars Hernquist for useful discussions. Y.N. acknowledges support from the ITC post-doctoral fellowship. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions; http://iaifi.org/). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under grant contract No. DE-SC0012567. The CAMELS project is supported by the Simons Foundation and NSF grant AST 2108078.

Appendix A
Impact of the Different Galaxy Formation Models in the Stellar Density Fields

In this section, we highlight the differences in the bias between stellar density fields and DM fields found in the different galaxy models.

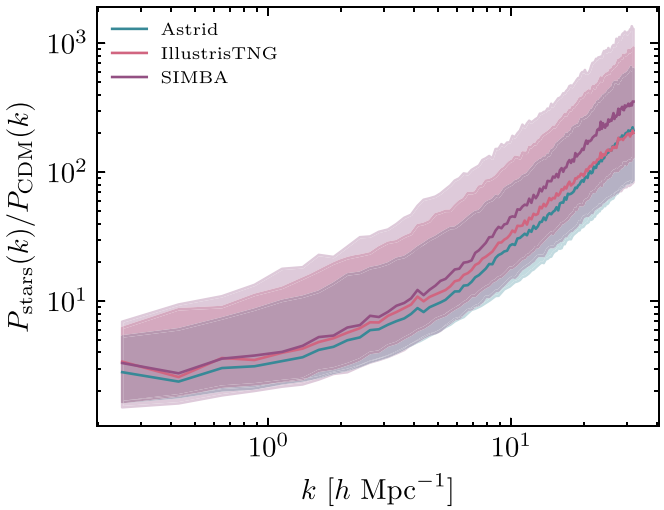

Figure 8. Ratios of star to DM fields power spectra from each data set, where the shaded region represents the range of ratios for the LH set and the solid line represents the ratio for the CV set. Note that we show the 10th–90th percentiles of the LH samples. In general, the different models show a consistent behavior as a function of scale.

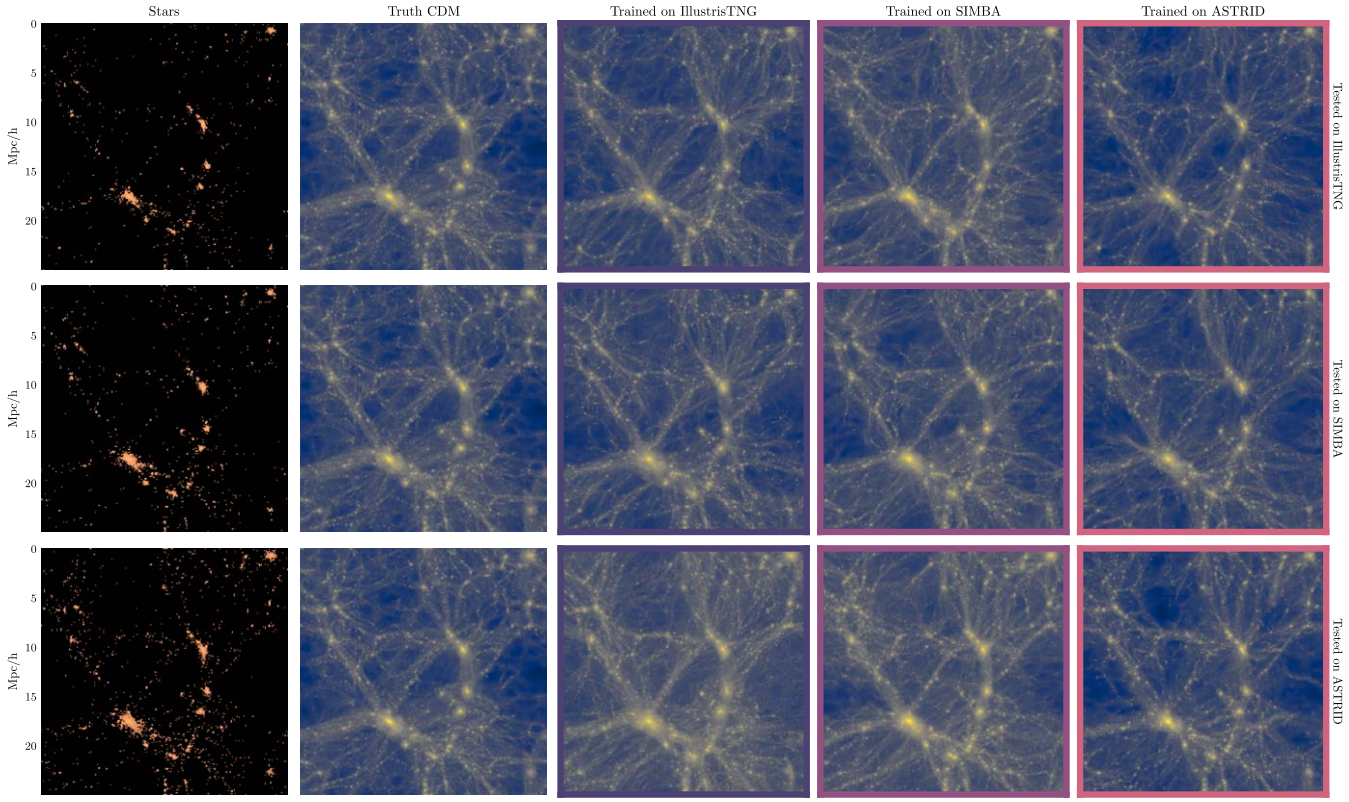

Figure 9. First column: input stellar fields for each of the simulations suites where each row shows the same-index images from the CAMELS CV sets (IllustrisTNG, SIMBA, ASTRID), whose parameter values are the same. Second column: corresponding DM field. Third to fifth columns: DM fields sampled from each model trained on IllustrisTNG, SIMBA, and ASTRID suites, respectively. The generated DM fields all correspond well to the true images visually.

In Figure 8, we show the ratio of star to DM power spectra of the different simulation suites. The solid lines show the CV set ratios of each simulation, while the shaded regions represent the 10th–90th percentiles of the LH samples. In general, the different models show a consistent behavior as a function of scale.

Moreover, Figure 9 depicts the different stellar fields for the same initial conditions and parameter values in the different simulation suites. It shows that although the DM density fields are practically the same in the different simulation suites, the stellar mass maps can look qualitatively different. We also show one sample from the diffusion model trained and tested in all possible combinations.

Appendix B
Ablation Study

Deep learning models consist of blocks and parameters that can be adjusted, and often these tweaks present a wide range of

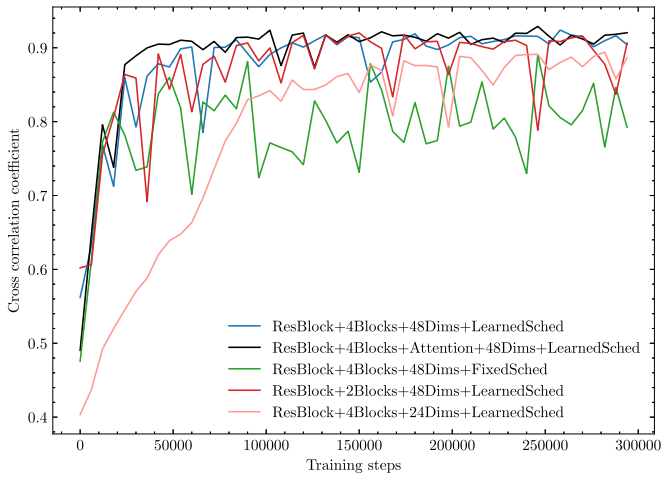

Figure 10. Average cross-correlations between true DM and 100 sampled DM fields at every 3000th checkpoint given the same conditioning sample and different diffusion model settings.

predictions. We therefore perform an ablation test to study the behavior of the diffusion model and examine how the removal or addition of components affects model performance. Our initial model consists of 4 convolution (ResNet) blocks, 48 dimensions of time embedding, and a learned linear schedule. From this model, we implement four other changes varied one at a time: adding an attention block to the bottom of the U-Net architecture, fixing the linear schedule, reducing the number of convolution blocks by half, and reducing the dimensions of time embedding by half.

Attention is a mechanism introduced by Vaswani et al. (2017) and is commonly used in deep learning models, namely, in natural language processing. In the context of images, attention lets the model highlight only the relevant features of the image. We conduct an ablation study with the addition of an attention block to the bottom of the U-Net architecture.

The convolution block is the fundamental block of the denoising model architecture, and four is the typical number used in U-Net as it is large enough to increase the number of feature channels sufficiently but small enough to keep the computational resources low. In this ablation test, we halve the number of blocks to observe how the reduced depth of feature channels may affect the NN's ability to learn the features.

In the forward diffusion process, the model adds noise to the input image in $T = 250$ steps according to a variance schedule. In our original model, we employ a schedule that changes its γ parameters in its function as it learns to predict better outputs. In this ablation study, we test its performance when we fix the schedule instead.

Time embeddings are how the NN shares its parameters across time as each denoising process occurs at each time step. In other words, t is encoded via time embeddings for the network to know what the current level of the noisy image is when processing at a given time. In our original model, we choose 48 as our time embedding dimensionality, and in this ablation study we perform one with 24.

We train each model using the IllustrisTNG simulation suite. We train across 300,000 steps, or equivalently 300 epochs with 1000 training steps at each epoch, and store checkpoints at every 3000 steps. We choose the metric for this test to be the average cross-correlations as these are the most fine-grained summary statistics, as discussed in Section 3.1.

Figure 10 shows the performance of each model across training steps. The model with an attention block added consistently yields the highest coefficients, while the one without follows very closely below. Halving the number of convolution blocks shows more fluctuations in their performance, which indicates that the original depth of the architecture is more optimal for it to be stable as the model learns. When using a fixed linear schedule, the model performs well at first, but since it does not learn the best γ values, its performance does not improve across training. The model with a half dimension of time embeddings significantly underperforms relative to the rest, though it catches up by the end of training.

We conclude from this study that the diffusion model with attention is very efficient in generating accurate DM fields. Since the addition of this block did not need additional computational resources from what we had allocated for all other models, we choose this model as our best and implement this architecture for further analysis with all simulation suites.

Appendix C
Detailed Analysis of Posterior Uncertainties

In Figure 11, we compare the ratio of posterior standard deviation to posterior mean as a function of the posterior mean for one input stellar mass sample from the ASTRID CV set.

In particular, Figure 11 shows the difference in behavior for pixels where the stellar mass is nonzero versus pixels with a zero stellar mass. When the stellar mass is low but nonzero, the posterior samples show a higher ratio of variance to the mean. This would correspond to small DM halos and filaments. On the other hand, when the stellar mass is high, the ratio of posterior variance to the mean is the lowest.

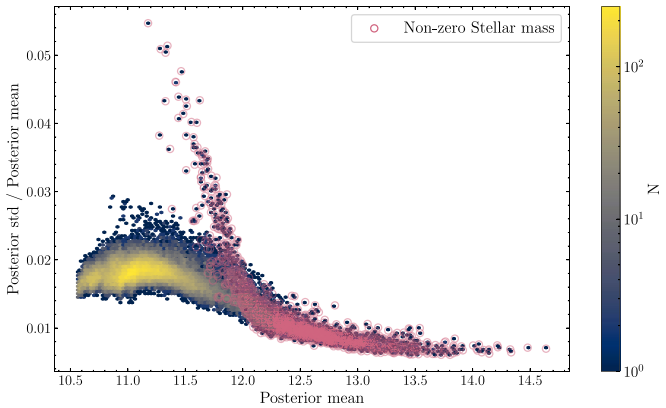

Figure 11. Joint probability density histogram of the posterior ratio of standard deviation to the mean and the posterior mean of the sampled CDM field. All pixels corresponding to locations with nonzero conditioning stellar mass are circled in pink.

Appendix D
Out-of-distribution Performance of the Parameter Inference Network

In Figure 12, we show the predictions of the parameter inference network trained on a subset of the Illustris-TNG LH set to infer Ω_m and σ_8 given a DM density field when tested on the DM density fields from the ASTRID LH set. In particular, Figure 12 demonstrates that the inference network produces biased values of σ_8.

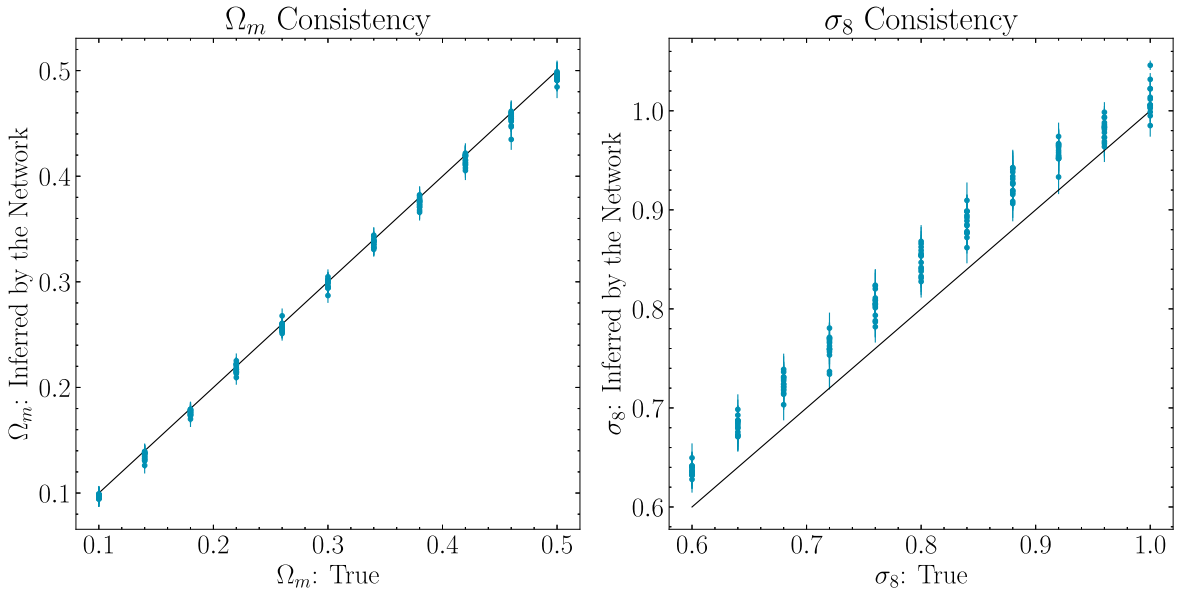

Figure 12. Performance of the parameter inference networks from Villaescusa-Navarro et al. (2021) on the true 1P DM fields for ASTRID. On the *x*-axis, we show the true Ω_m and σ_8 values for each simulation. The *y*-axis shows the mean and variance of the posterior predictions for the cosmological parameters given a DM density field from the ASTRID LH simulations.

ORCID iDs

Victoria Ono https://orcid.org/0009-0002-7522-9566
Core Francisco Park https://orcid.org/0000-0002-9542-2913
Nayantara Mudur https://orcid.org/0000-0001-5139-612X
Yueying Ni https://orcid.org/0000-0001-7899-7195
Carolina Cuesta-Lazaro https://orcid.org/0000-0002-6069-2999
Francisco Villaescusa-Navarro https://orcid.org/0000-0002-4816-0455

References

Cuesta-Lazaro, C., & Mishra-Sharma, S. 2023, PhRvD, 109, 123531
de Santi, N. S. M., Shao, H., Villaescusa-Navarro, F., et al. 2023, ApJ, 952, 69
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Desjacques, V., Jeong, D., & Schmidt, F. 2018, PhR, 733, 1
Falcon, W., Borovec, J., & Wälchli, A. 2020, PyTorch Lightning/pytorch-lightning v0.7.6, Zenodo, doi:10.5281/zenodo.3828935
Hahn, C., Eickenberg, M., Ho, S., et al. 2023, PNAS, 120, e2218810120
He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv:1512.03385
Hong, S. E., Jeong, D., Hwang, H. S., & Kim, J. 2021, ApJ, 913, 76
Ivanov, M. M. 2021, PhRvD, 104, 103514
Kingma, D., Salimans, T., Poole, B., & Ho, J. 2021, in 35th Conf. on Neural Information Processing Systems: Advances in Neural Information Processing Systems 34, ed. M. Ranzato et al. (Red Hook, NY: Curran Associates Inc.), 21696
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Li, Y., Hu, W., & Takada, M. 2014, PhRvD, 89, 083519
Loshchilov, I., & Hutter, F. 2017, arXiv:1711.05101
Loshchilov, I., & Hutter, F. 2016, arXiv:1608.03983
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Mudur, N., & Finkbeiner, D. P. 2022, arXiv:2211.12444
Nelson, D., Springel, V., Pillepich, A., et al. 2021, ComAC, 6, 2
Nguyen, N.-M., Schmidt, F., Tucci, B., Reinecke, M., & Kostić, A. 2024, arXiv:2403.03220
Ni, Y., Genel, S., Anglés-Alcázar, D., et al. 2023, APJ, 959, 136
Park, C. F., Ono, V., Mudur, N., Ni, Y., & Cuesta-Lazaro, C. 2023, arXiv:2311.08558
Pillepich, A., Nelson, D., Hernquist, L., et al. 2017, MNRAS, 475, 648
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. 2021, arXiv:2112.10752
Ronneberger, O., Fischer, P., & Brox, T. 2015, in Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, ed. N. Navab et al. (Cham: Springer), 234
Senatore, L. 2015, JCAP, 2015, 007
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757
Springel, V., Pakmor, R., Pillepich, A., et al. 2018, MNRAS, 475, 676
Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, arXiv:1706.03762
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021, ApJ, 915, 71
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021, arXiv:2109.09747
Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. 2022, ApJS, 259, 61
Villanueva-Domingo, P., & Villaescusa-Navarro, F. 2021, ApJ, 907, 44
Vogelsberger, M., Marinacci, F., Torrey, P., & Puchwein, E. 2020, NatRP, 2, 42
Wu, Y., & He, K. 2018, arXiv:1803.08494