

---

# Cosmological Field Emulation and Parameter Inference with Diffusion Models

---

**Nayantara Mudur**

Harvard University  
Cambridge, MA, 02138, USA  
nmudur@g.harvard.edu

**Carolina Cuesta-Lazaro**

The NSF AI Institute for Artificial Intelligence and Fundamental Interactions  
Massachusetts Institute of Technology, Cambridge, MA, 02139, USA  
cuestalz@mit.edu

**Douglas P. Finkbeiner**

Harvard University  
Cambridge, MA, 02138, USA  
dfinkbeiner@cfa.harvard.edu

## Abstract

Cosmological simulations play a crucial role in elucidating the effect of physical parameters on the statistics of fields and on constraining parameters given information on density fields. We leverage diffusion generative models to address two tasks of importance to cosmology – as an emulator for cold dark matter density fields conditional on input cosmological parameters  $\Omega_m$  and  $\sigma_8$ , and as a parameter inference model that can return constraints on the cosmological parameters of an input field. We show that the model is able to generate fields with power spectra that are consistent with those of the simulated target distribution, and capture the subtle effect of each parameter on modulations in the power spectrum. We additionally explore their utility as parameter inference models and find that we can obtain tight constraints on cosmological parameters.

## 1 Introduction

Cosmological simulations are expensive to run, and can only be generated for a limited set of initial conditions and points in parameter space. This has led to the need for emulators or surrogate models [Heitmann et al., 2009, Mustafa et al., 2019] that can learn to model the distribution of fields or summary statistics of importance to cosmology. A closely interlinked thrust of modern cosmology has been the search for statistics that can yield optimal constraints on cosmological parameters [Valogiannis and Dvorkin, 2022, Dai and Seljak, 2023]. Diffusion or score-based generative models [Song et al., 2020] are generative models that involve a forward diffusion (noising) process. A neural network is then used to learn the denoising or generative process that maps samples from the standard normal to samples from the target distribution. The denoising diffusion probabilistic model (DDPM) [Ho et al., 2020] formulation consists of a variance schedule  $\beta_t$  over a fixed number of time steps,  $T$ , that controls the incremental noise added to the image, or the ‘speed’ at which the target distribution is noised, and a score model, that is used to parameterize the reverse process. The model can also be conditional on either a discrete input label or a parameter vector. In this work, we apply diffusion generative models to emulate fields from cosmological simulations and show that they can

capture the dependence of summary statistics on input cosmology, and can provide tight constraints on cosmological parameters.

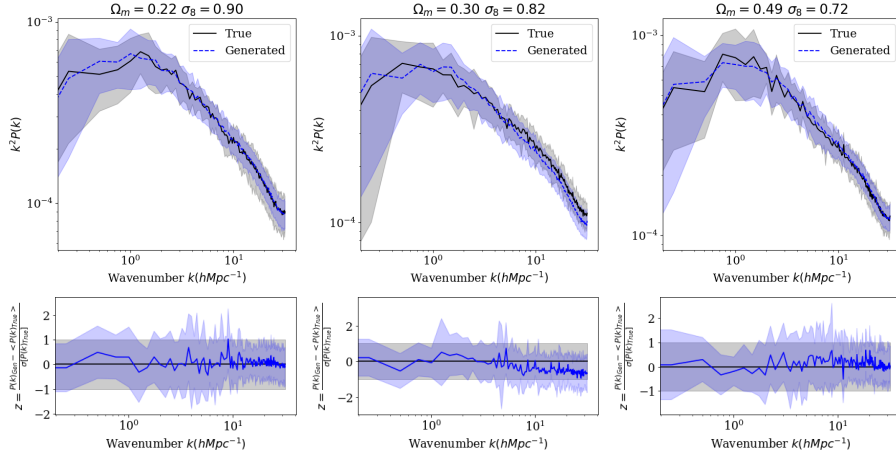


Figure 1:  $k^2 P(k)$  (first row) and the  $z$ -score (second row) for 3 different validation parameters.

## 2 Datasets, Architecture and Training

**Datasets** We used the IllustrisTNG [Nelson et al., 2019, Pillepich et al., 2018] Cold Dark Matter density fields at  $z = 0$  from the CAMELS Multifield Dataset (CMD) [Villaescusa-Navarro et al., 2022, 2021a]. The dataset consists of 15 two-dimensional fields each for 1000 different points in parameter space. The parameter vector has 2 cosmological ( $\Omega_m$  and  $\sigma_8$ ) and 4 astrophysical parameters. The fields span  $25h^{-1}$  Mpc on each side. We work with the log (base 10) of these fields and randomly rotate or flip the field to account for invariance under rotations or parity.

**Diffusion Model Setup** The forward diffusion process follows a variance schedule  $\{\beta_t\}$ , and  $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$ , where  $\bar{\alpha}_t = \prod_{t'=1}^t (1 - \beta_{t'})$ . The diffusion model architecture is similar to the U-Net [Ronneberger et al., 2015] used in Ho et al. [2020], has 4 down and up-sampling blocks consisting of 2 ResNet blocks [Zagoruyko and Komodakis, 2016], group-normalization [Wu and He, 2018], and attention [Vaswani et al., 2017, Shen et al., 2021]. We use circular convolutions in the downsampling layers since the input fields have periodic boundary conditions. The model is conditional on  $\Omega_m$  and  $\sigma_8$ . Each parameter is normalized to lie between  $[0, 1]$  with respect to its range,  $\Omega_m \in [0.1, 0.5]$ ,  $\sigma_8 \in [0.6, 1.0]$ . The model has a multilayer perceptron that is used to transform the parameters into a space with the same dimension as the time embedding, and each ResNet block additionally has an MLP that takes cosmology as an input. We used a non-linear variance schedule with 1000 timesteps. During training, for each image  $\mathbf{x}_0$ , a timestep is sampled uniformly along with a noise pattern  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ . The loss function that is minimized is  $L_{t-1} = \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t, \vec{y})\|^2$ , where  $\vec{y}$  is the parameter vector. We used the Weights and Biases framework [Biewald, 2020] for our experiments.

**Training** We first train the conditional diffusion model on downsampled  $64 \times 64$  images for 60000 iterations. To train the conditional diffusion model on  $256 \times 256$  images, we initialize the architecture with the weights of the  $64 \times 64$  model after 60000 iterations, and then train the model for over 500000 iterations. In our experiments, initializing the  $256 \times 256$  model with the weights of the  $64 \times 64$  model appeared to lead to faster convergence. The diffusion model loss is not informative of sample quality, and we need an alternative metric to assess the quality of the generated samples Theis et al. [2015]. We sample 50 fields for 10 different validation parameters for 4 different checkpoints – corresponding to 200k, 220k, 240k and 260k iterations. For each checkpoint we compute the reduced chi-squared statistic of the power spectrum of each generated field,  $s$ , with respect to the power spectra of the 15 true fields from the dataset, corresponding to that parameter:  $\chi_r^2(s) = \frac{1}{|k|-1} \sum_k \frac{(P(k)_s - \langle P(k)_{\text{TRUE}} \rangle)^2}{\sigma[P(k)_{\text{TRUE}}]^2}$ . We then compute the mean of these values across all parameters and sampled fields. The checkpoint

corresponding to the 260k iteration had the lowest value, corresponding to 1.30. To put this number in perspective, we can examine the effect of cosmic variance on this metric using a leave-one-out cross-validation approach, by computing the reduced chi-squared statistic of each sample of a *true* field, using the 14 other true fields corresponding to the same parameter as the reference distribution. The mean of this value across the 10 parameters is 1.27. This is consistent with the value of 1.27 that is obtained if the reduced chi-squared statistic is computed in a leave-one-out fashion using 15 observations of independently distributed Gaussians of length 128 (the number of bins in the power spectrum). We use the 260k checkpoint for our analysis.

### 3 Summary Statistics

We examine the power spectra for 3 different validation parameters for 15 true fields and 15 generated fields from the diffusion model, along with the  $z$ -scores in each  $k$  bin in Figure 1. All statistics in this section are computed on the log of the true fields and the generated fields. In Figure 2, we generated ‘1P’ sets and examined whether the effect of changing each parameter, while keeping the others constant is the same as is observed in the 1P CAMELS suite. Increasing  $\Omega_m$  enhances the power spectrum at all scales in the generated fields (center) and affects the pixel values. Changes in  $\sigma_8$  affect only scales larger than a few  $h^{-1}$  Mpc. The one sigma envelope for the ratios of the modulations for the true fields appears to be slightly larger than that for the generated fields.

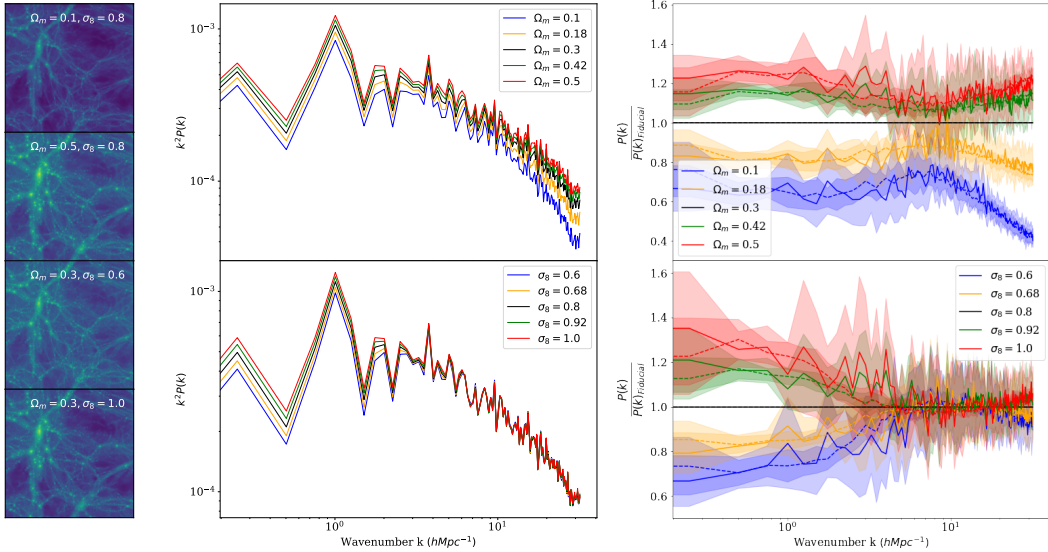


Figure 2: Generated ‘1P’ fields. **Left column:** Generated fields corresponding to the extreme values of each parameter for a single seed, with the other value held fixed at the fiducial value (0.3 for  $\Omega_m$  and 0.8 for  $\sigma_8$ ). **Middle column:** Power spectra of the generated fields for the same seed, for different values of each parameter, holding the other fixed. **Right column:** Mean and standard deviation for the ratio of the power spectra at the modified parameter value to the power spectra for the field at the fiducial parameter value (black) for 15 slices from the CAMELS dataset (solid) and 15 seeds for the generated fields from the diffusion model (dashed).

### 4 Parameter Inference

Diffusion models allow an evaluation of the lower bound on the log likelihood, the variational lower bound (VLB).  $L_{vlb} = L_0 + L_1 \dots L_{T-1} + L_T = -\log p_\theta(x_0|x_1) + \sum_{t>1} D_{KL}[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)] + D_{KL}[q(x_T|x_0)||p(x_T)]$  where  $p_\theta$  are the learned reverse distributions, and  $q$  are the forward (analytical) distributions. Since the diffusion model is conditioned on an input parameter, we can derive an upper bound on the negative log likelihood, conditional on an input parameter. One can thus investigate whether the variational lower bound terms of a trained diffusion model can be interpreted as a statistic that is sensitive to the parameter corresponding

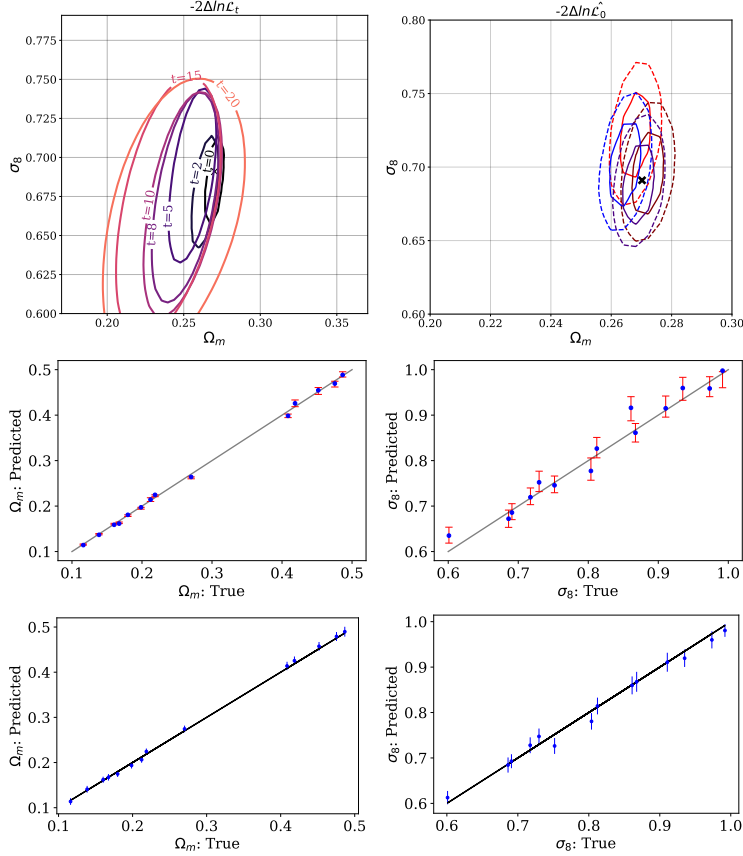


Figure 3: **First Row, Left:** One sigma contours for  $-2\Delta\ln\hat{\mathcal{L}}$  contribution for 7 different timesteps. **First Row, Right:** One (solid) and two (dashed) sigma contours for the  $-2\Delta\ln\hat{\mathcal{L}}$  term for the 0th timestep, for 2 different seeds and 2 different samples of the input fields for a specific parameter. The cross demarcates the true parameter corresponding to the input fields. **Second Row:** Predicted parameter and ground truth parameter for 14 different validation parameters. The error bars correspond to the 68% interval for the marginal probability distributions for each parameter for each input field using the  $-2\Delta\ln\hat{\mathcal{L}}_0$  term. **Third Row:** Predicted parameter and ground truth parameter for 14 different validation parameters using the parameter inference networks in Villaescusa-Navarro et al. [2021b].

to a given input field. For an input field  $x_0$ , we evaluate the  $L_t(x_0|\vec{\theta}_{\text{EVAL}})$  terms over a grid in  $[\Omega_m, \sigma_8]$ , centered on the value of the true field  $\vec{\theta}_{\text{TRUE}}$ . The grid spans  $\max(\Omega_{m,\text{TRUE}} - 0.1, 0.1)$  and  $\min(\Omega_{m,\text{TRUE}} + 0.1, 0.5)$  in  $\Omega_m$  and likewise for  $\sigma_8$ , with 50 points in both dimensions, resulting in an evaluation grid with 2500 points. To map each term's contribution to a chi-squared distribution, we first subtract the minimum value of  $L_t(x_0|\vec{\theta}_{\text{EVAL}})$  on the grid, and multiply it by 2 to yield  $-2\Delta\ln\hat{\mathcal{L}}_t$  for each  $t$ . We can then identify the 1, 2, and 3 sigma contours corresponding to this estimated chi-squared distribution with 2 degrees of freedom. We plot the contours for seven such timesteps in the top left panel of Figure 3. Since larger time steps correspond to progressively noisier images, the first few terms possess the most discriminatory power. Since the change in  $L_{\text{vib}}(x_0|\vec{\theta}_{\text{EVAL}})$  from one value of  $\vec{\theta}_{\text{EVAL}}$  to another is dominated by the  $L_0$  term, we approximate the likelihood ratio by the contribution arising from the ratio of the first term  $L_0$ . We defer a more rigorous examination of the optimum subset of timesteps to use to optimize the tradeoff between faster computation and higher precision to future work. Next, we test whether the approximation to  $-2\Delta\ln\hat{\mathcal{L}}(\vec{\theta}_{\text{EVAL}}|x_0)$  computed using only  $L_0$  is minimized at  $\vec{\theta}_{\text{TRUE}}$ , the true value of the parameter corresponding to the input field  $x_0$ . There are two sources of stochasticity that contribute to this test (see top right, Figure 3): the seed used to sample  $x_{t=1}$  in the  $L_0$  term, and the choice of input field sample from the true

distribution for the same  $\vec{\theta}_{\text{TRUE}}$  (cosmic variance). In the second row, we plot the true and predicted cosmological parameters for 14 different parameters from the validation set for a single sample and seed for each parameter using only the  $-2\Delta\ln\hat{\mathcal{L}}_0$  term. We convert  $-2\Delta\ln\hat{\mathcal{L}}_0$  to a likelihood and sum over each axis to derive a marginal likelihood for each parameter on the grid. The ‘predicted’ parameter is the parameter at which the marginal probability distribution is maximized on the 1D grid. The one-dimensional error bars, are obtained by finding the 68% confidence interval for each marginal. Note, that since the marginals do not account for the covariance,  $\vec{\theta}_{\text{TRUE}}$  might lie within the one-dimensional 1 sigma interval without lying within the two-dimensional 1 sigma contour. The constraints on  $\Omega_m$  are much tighter than those on  $\sigma_8$ . This is consistent with and comparable to the performance of the parameter inference networks in Villaescusa-Navarro et al. [2021b] (third row). The predicted  $\vec{\theta}_{\text{TRUE}}$  is close to the true value of  $\vec{\theta}_{\text{TRUE}}$  over a broad range of parameters. Note, the 1D error bars in Villaescusa-Navarro et al. [2021b] are not derived from a two dimensional approximate likelihood, but are derived from a neural network that is trained to return the mean and the standard deviation of each parameter. Since the diffusion model’s noise prediction loss is related to the reweighted VLB Kingma and Gao [2023], the terms of the VLB computed using the conditional diffusion model encode dependencies on the cosmological parameters. The error bars on  $\Omega_m$  often do not account for the error in the prediction, and we intend to explore whether including more VLB terms yields better calibrated error bars. It would also be useful to plug the conditional diffusion model-based approximate likelihood ratios into an MCMC sampler Hermans et al. [2020] and compare the constraints we obtain to those obtained from canonical summary statistics such as the power spectrum.

## 5 Conclusion

In this work, we deployed a diffusion generative model as an emulator for log cold dark matter density fields, and as a parameter inference model that can yield tight constraints on cosmological parameters. Future work could be focused on finding ways to accelerate the generation process. We also intend to examine the statistics of the fields in regular (non-log) space, convergence with respect to the number of seeds used to estimate the likelihood-based constraints and the subset of terms needed for the inference constraints to enhance the calibration of our parameter inference step. It would be interesting to compare the performance of the parameter inference step on the true fields vs the generated fields, fields from the CAMELS simulation suite with a different choice of astrophysical feedback, such as SIMBA, and the effect of adding distortions to the image. Parameter inference approaches that are able to marginalize over astrophysical feedback and observational noise are desirable.

## 6 Acknowledgements

We thank Yueying Ni, Core Francisco Park, Shuchin Aeron, Andrew K. Saydjari, and Justina R. Yang for helpful discussions. This work was supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

## References

- Katrin Heitmann, David Higdon, Martin White, Salman Habib, Brian J. Williams, and Christian Wagner. The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum. *Astrophys. J.*, 705:156–174, 2009. doi: 10.1088/0004-637X/705/1/156.
- Mustafa Mustafa, Deborah Bard, Wahid Bhimji, Zarija Lukić, Rami Al-Rfou, and Jan M Kratochvil. Cosmogon: creating high-fidelity weak lensing convergence maps using generative adversarial networks. *Computational Astrophysics and Cosmology*, 6(1):1–13, 2019.
- Georgios Valogiannis and Cora Dvorkin. Going beyond the galaxy power spectrum: An analysis of boss data with wavelet scattering transforms. *Physical Review D*, 106(10):103509, 2022.
- Biwei Dai and Uros Seljak. Multiscale flow for robust and optimal cosmological analysis. *arXiv preprint arXiv:2306.04689*, 2023.

- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez-Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, et al. The illustrixtng simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1): 1–29, 2019.
- Annalisa Pillepich, Volker Springel, Dylan Nelson, Shy Genel, Jill Naiman, Rüdiger Pakmor, Lars Hernquist, Paul Torrey, Mark Vogelsberger, Rainer Weinberger, et al. Simulating galaxy formation with the illustrixtng model. *Monthly Notices of the Royal Astronomical Society*, 473(3):4077–4106, 2018.
- Francisco Villaescusa-Navarro, Shy Genel, Daniel Anglés-Alcázar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, et al. The camels multifield data set: Learning the universe’s fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2):61, 2022.
- Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N Spergel, Rachel S Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, et al. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, 2021a.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.
- Lukas Biewald. Experiment tracking with weights and biases, 2020. URL <https://www.wandb.com/>. Software available from wandb.com.
- Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.
- Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N Spergel, Yin Li, Benjamin Wandelt, Andrina Nicola, Leander Thiele, Sultan Hassan, Jose Manuel Zorrilla Matilla, et al. Multifield cosmology with artificial intelligence. *arXiv preprint arXiv:2109.09747*, 2021b.
- Diederik P Kingma and Ruiqi Gao. Understanding the diffusion objective as a weighted integral of elbos. *arXiv preprint arXiv:2303.00848*, 2023.
- Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR, 2020.

## 7 Appendix

### 7.1 Additional Summary Statistics

We examine the consistency of the pixel histograms of the fields for the 3 validation parameters in Figure 1. The model appears to be slightly biased toward producing fields with higher density for the same parameter. We hope to find ways to mitigate this artefact in future work. In Figure 5, we plot the histograms of the distributions of the reduced chi-squared statistics computed as described in Section 2. The plots on the left and the center serve as an ‘apples-to-apples’ comparison since they use the same reference distributions and are tested on the same number of fields (the remaining true field or a generated field). While the plot with the reduced chi-squared statistic for 500 fields has some outliers, we find that the three distributions are in good agreement with each other.

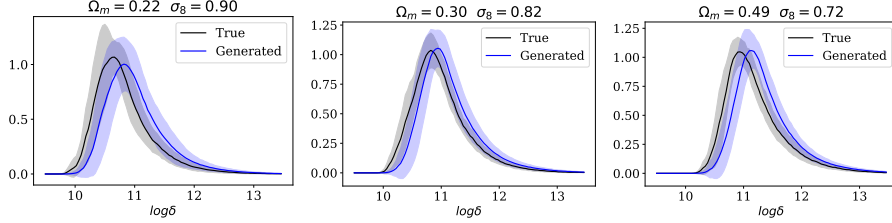


Figure 4: Density histograms plotting the mean and the standard deviation envelope for 15 true fields and 50 generated fields for each parameter.

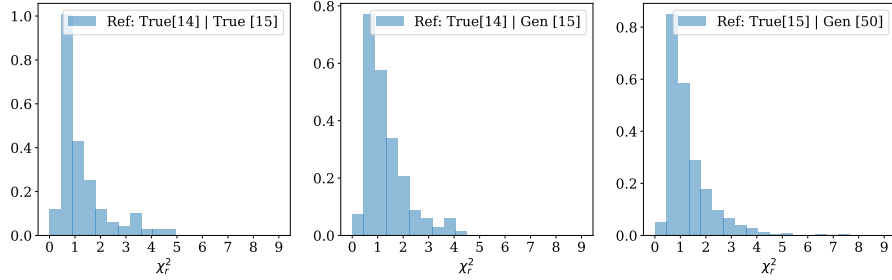


Figure 5: Distribution of the reduced chi-squared statistic for the power spectra of the generated fields relative to the true fields belonging to the same parameter for 10 different validation parameters. *Left, Center*: ‘Leave-one-out’ method: 14 true fields’ power spectra serve as the reference distribution and we test on the power spectra of either the remaining true field (left) or a generated field (center). We use 15 test fields per parameter for both cases and have 150 datapoints in each. *Right*: All 15 true fields’ power spectra serve as the reference distribution and we test on all 50 generated fields. Since there are 50 fields and 10 parameters, we have 500 datapoints in this plot.

### 7.2 Parameter Inference

$$\mathbb{E}_q[-\log p_{\text{NN},\theta_{\text{EVAL}}}(x_0)] \leq \mathbb{E}_q[\text{D}_{\text{KL}}[q(x_{\text{T}}|x_0)||p(x_{\text{T}})]] + \quad (1)$$

$$\sum_{t>1} \text{D}_{\text{KL}}[q(x_{t-1}|x_t, x_0)||p_{\text{NN},\theta_{\text{EVAL}}}(x_{t-1}|x_t) - \log p_{\text{NN},\theta_{\text{EVAL}}}(x_0|x_1)]$$

$$\text{For } t=0, -2\Delta \ln \hat{\mathcal{L}}_0(\theta_{\text{EVAL}}|x_0) \simeq 2[L_0 - \text{argmin}_{\theta_{\text{EVAL}}} L_0], \text{ where } L_0 = -\log p_{\text{NN},\theta_{\text{EVAL}}}(x_0|x_1)$$

$$\text{For } t \in [1, T-1], -2\Delta \ln \hat{\mathcal{L}}_t(\theta_{\text{EVAL}}|x_0) \simeq 2[L_t - \text{argmin}_{\theta_{\text{EVAL}}} L_t] \quad (2)$$

$$\text{where } L_t = \text{D}_{\text{KL}}[q(x_{t-1}|x_t, x_0)||p_{\text{NN},\theta_{\text{EVAL}}}(x_{t-1}|x_t)]$$

As next steps, we explore a more precise approximation to the VLB using the sum of a subset of the  $L_t$  terms. Here, the minimum value over the grid is computed over the sum of the terms.