Inpainting Galaxy Counts onto N-Body Simulations over Multiple Cosmologies and Astrophysics

Antoine Bourdin [1 2 3] **Ronan Legin** [2 3 4] **Matthew Ho** [5] **Alexandre Adam** [2 3 4] **Yashar Hezaveh** [2 3 4 6 7 8]
Laurence Perreault-Levasseur [2 3 4 6 7 8]

Abstract

Cosmological hydrodynamical simulations, while the current state-of-the art methodology for generating theoretical predictions for the large scale structures of the Universe, are among the most expensive simulation tools, requiring upwards of 100 millions CPU hours per simulation. N-body simulations, which exclusively model dark matter and its purely gravitational interactions, represent a less resource-intensive alternative, however, they do not model galaxies, and as such cannot directly be compared to observations. In this study, we use conditional score-based models to learn a mapping from N-body to hydrodynamical simulations, specifically from dark matter density fields to the observable distribution of galaxies. We demonstrate that our model is capable of generating galaxy fields statistically consistent with hydrodynamical simulations at a fraction of the computational cost, and demonstrate our emulator is significantly more precise than traditional emulators over the scales 0.36 h Mpc^{-1} $\leq$ k $\leq$ 3.88 h Mpc^{-1}.

1. Introduction

One of the most powerful tools in cosmology are hydrodynamical (hydro-) simulations, which are able to evolve the universe from billions of years into the past until present day (e.g. Nelson et al., 2019; Davé et al., 2019; Bird et al., 2022). These simulations make detailed and reliable predictions for the distribution of matter in the universe by carefully modeling the complex interactions of gravity, mass, thermodynamics, and electrodynamics. They have developed into a cornerstone of modern cosmology, creating a gateway connecting observations of the night sky with physical understandings of dark matter (Bozorgnia & Bertone, 2017), dark energy (Baldi et al., 2010), galaxies (Vogelsberger et al., 2020) and black holes (Chen et al., 2022).

These powerful simulations come at a cost: they are extremely computationally expensive. High-resolution hydro-simulations can require hundreds of millions of CPU-hours to properly resolve galactic astrophysics (Vogelsberger et al., 2014). This computational cost is due to the $O(n^2)$ scaling of computing thermodynamic interactions between dark matter, gas, and star particles and is a major barrier for modern cosmology. Furthermore, recent works attempt to run not only one hydro-simulation, but suites of thousands (e.g. Villaescusa-Navarro et al., 2023; Schaye et al., 2023), with each simulation varying the configurations of cosmological models and galaxy astrophysics. These suites are essential to understand the effect of (latent) physical model parameters on astronomical observables, hence making it possible to infer the value of these latent parameters for our Universe from survey data, with traditional method but also using machine learning (e.g. de Santi et al., 2023; Kugel et al., 2023; Ho et al., 2023; 2024). With the increasing quality and volumes of survey data, such suites of hydro-simulations are projected to drive the next-generation of computational cosmology (National Academies of Sciences, Engineering, and Medicine, 2023), and it is therefore paramount to develop faster methods for running them which exceed the traditional scaling laws.

A cheaper alternative to hydro-simulations are N-body simulations (e.g. Heitmann et al., 2019; Maksimova et al., 2021; Villaescusa-Navarro et al., 2020), which only seek to model interactions between collision-less dark matter particles. Ignoring thermodynamic interactions reduces the computational scaling of N-body simulations to $O(n \log n)$, a marked improvement over hydro-simulations. N-body

[1]Department of Physics, McGill University, Montréal, Canada [2]Ciela Institute, Montréal, Canada [3]Mila - Quebec Artificial Intelligence Institute, Montréal, Canada [4]Department of Physics, Université de Montréal, Montréal, Canada [5]CNRS & Sorbonne Université, Institut d'Astrophysique de Paris (IAP), Paris, France [6]Center for Computational Astrophysics, Flatiron Institute, New York, USA [7]Perimeter Institute for Theoretical Physics, Waterloo, Canada [8]Trottier Space Institute, McGill University, Montréal, Canada. Correspondence to: Antoine Bourdin <antoine.bourdin@mail.mcgill.ca>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

simulations are accurate approximations for the spatial distribution of all matter, as dark matter is believed to make up $\sim 85\%$ of the total matter budget in the universe. However, as N-body simulations do not directly simulate stars and galaxies, we must rely on a plethora of simplifying assumptions when comparing them to real observations (Desjacques et al., 2018). The parameters of these assumptions are particularly challenging to constrain for populations of smaller galaxies which dominate the statistics of observational surveys (e.g. Richardson et al., 2012; Yuan et al., 2022; Garcia-Quintero et al., 2024). In these high-resolution regimes, the use of hydro-simulations, or their respective emulators, is crucial.

In this work, we develop an emulator for high-resolution hydro-simulations by populating N-body simulations with galaxies. We make use of a conditional score-based diffusion model (Song et al., 2021; Legin et al., 2023) that generates samples containing the number of galaxies at different grid positions given the mass of matter evolved in an N-body simulation over a pixelated grid.

Our model is trained on the 3D CAMELS-Astrid simulation from the LH set (Villaescusa-Navarro et al., 2022), which consists of one thousand pairs of N-body and hydro-simulations. Each pair of simulations has a different configuration of cosmological and astrophysical parameters, varying the total matter density (Ω_m), the amplitude of matter density fluctuations (σ_8), the energy of galactic winds ($A_{\rm SN1}$), the wind speed of galactic winds ($A_{\rm SN2}$), the rate of energy injection from active galactic nuclei ($A_{\rm AGN1}$), and the radio mode threshold for active galactic nuclei ($A_{\rm AGN2}$). Each of these parameters has a strong, non-linear effect on the spatial distribution of galaxies in the universe.

Using a diffusion model with this dataset allows us to learn the distribution of galaxy counts over multiple cosmologies. This means that, given an N-body simulation at a fixed cosmology, we can generate in parallel 100 different realizations of galaxy counts which match the summary statistics of a hydro-simulation in 20 minutes. Meanwhile, obtaining an analogous result with traditional methods would entail running one hundred hydro-simulations, each requiring millions of CPU hours.

Furthermore, we show that our model is able to correctly predict galaxy counts on cosmological and astrophysical parameters that were not seen during training, enabling us to explore how the parameter space of configurations affects galaxy clustering in a time-efficient manner. To the best of our knowledge, this is the first work connecting probabilistically N-body simulations to galaxies with the degree of accuracy of a hydro-simulations over a wide range of cosmological and astrophysical parameter values.

2. Related Works

Within this section we briefly discuss alternative works aiming to connect N-body simulation to baryonic properties. In section 2.1, we present the most commonly used closed-form parametric model used in cosmology while in section 2.2 we highlight other recent approaches which use machine learning. We highlight some of the advantages and limitations of each model.

2.1. Halo Occupancy Distribution

To circumvent the need for expensive hydrodynamical simulations, Halo Occupancy Distribution (HOD) models are often used to connect N-body simulations to galaxies (Zheng et al., 2005; 2007). As the name suggests, these models identify halos which are clustering of dark matter over a certain length scale. Halos create gravitational potential wells which attract baryons and becomes locations where galaxies form.

Once dark matter halos are identified in an N-body simulation, individual halos are assigned a probability of hosting a galaxy given their mass. The HOD model separates galaxies into two categories: a central galaxy sitting at the minima of the gravitational well and satellite galaxies which are progressively added and that orbit around the central galaxy. The expected value for the number of central ($N_{\rm cen}$) and satellite ($N_{\rm sat}$) galaxy as a function of halo mass (M) is described by,

$$\langle N_{\rm cen}(M)\rangle = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log M - \log M_{\rm min}}{\sigma_{\rm logM}}\right)\right]$$
$$\langle N_{\rm sat}(M)\rangle = \langle N_{\rm cen}(M)\rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}. \tag{1}$$

Here, erf is the Gaussian error function. The parameter $M_{\rm min}$ is the halo mass at which there is a 0.5 expected value of hosting a central galaxy. On the other hand, $\sigma_{\rm logM}$ controls how fast this expectation grows to one for different halo masses. Similarly, M_0 determines the minimum halo mass to have a non-zero expectation of hosting a satellite galaxy. The parameter M_1 and α control how fast this expectation grows. We note that here the expectation is not bounded by one as we can have multiple satellite galaxies within a single halo.

An HOD works by identifying all of the halos in a N-body along with their position and mass. It parses through each individual halo and computes the expected number of galaxies based on equation (1). For central satellites, a Bernoulli distribution is sampled with the expectation value set by $\langle N_{\rm cen}(M)\rangle$ to determine whether that halo hosts a central satellite. If so, a Poisson distribution with the expectation $\langle N_{\rm sat}(M)\rangle$ is sampled to populate the halo with satellites.

Because of their simplicity, HOD models are very inexpensive, and also have the notable advantage that they produce population of galaxies in a probabilistic manner. However, previous works have shown that simple HODs broadly marginalize over complex baryonic physics, thus lacking sufficient precision to match the data quality that is expected from upcoming observational surveys (Hadzhiyska et al., 2020). Furthermore, the HOD model described above ignores the effect of environment, which is known to significantly impact the process of galaxy formation. Finally, the parametrization in equation (1) does not include cosmological parameters.

2.2. Alternative ML Approaches

In prior work, machine learning techniques have been employed to estimate galaxy counts from N-body only simulations. For instance, (Zhang et al., 2019) utilized convolutional neural networks to map from N-body simulations to galaxy counts, achieving superior accuracy compared to the traditional Halo Occupation Distribution (HOD). It is presumed these improved results were due to the fact that CNNs capture local information, hence, focusing on environmental factors instead of only considering halo mass as is done in an HOD. Nevertheless, a significant limitation of this approach is the deterministic nature of the model's predictions concerning galaxy numbers. Such deterministic predictions fail to encapsulate the inherent stochasticity of the task at hand, since the information from dark matter N-body simulations is insufficient to exactly predict galaxy counts from hydrodynamical simulations.

Score-based generative models have shown significant promise in recent field-level inference tasks, where samples can be generated from learned probability distributions over cosmological fields (Legin et al., 2023). These models can address the limitations inherent in deterministic approaches as noted by (Zhang et al., 2019), while simultaneously providing a more realistic representation of galaxy counts compared to the Halo Occupation Distribution (HOD) model. Similarly, Ono et al. (2024) used score-based generative models to infer the dark matter distribution from the galaxy field, using two-dimensional slices of hydro-simulations. While this approach was applied to a task opposite to that of the generative model under consideration here, this work demonstrated that score-based generation could effectively capture the dark-matter-galaxy connection.

Motivated by these previous points, we explore the use of score-based generative modeling to develop a more accurate model for predicting galaxy counts from dark matter N-body simulations. Furthermore, we demonstrate that the score-based model can also learn to implicitly marginalize over cosmological parameters and improve accuracy over multiple cosmologies. A similar methodology was taken by Cuesta-Lazaro & Mishra-Sharma (2023) which used diffusive graph neural networks to model the spatial distribution of halos in dark-matter-only simulations. While aptly capturing the uncertainty of generative modeling, this approach was limited to modeling the dark-matter-only component of simulations and also required complex graph representations, which greatly limited the generative capacity to only 5,000 halos under GPU memory limitations. Our chosen approach using convolutional architectures allows for simple factorization of the generative distribution over the box volume, allowing for effective 'outpainting' to arbitrarily large volumes (e.g. Rouhiainen et al., 2023).

In the following section, we provide a concise overview of score-based generative models and their application in sampling the probability distribution $p(\boldsymbol{x}|\boldsymbol{y})$ of plausible realizations of galaxy count fields $\boldsymbol{x}$ conditioned on dark matter N-body simulations $\boldsymbol{y}$.

3. Score-Based Models

Score-based generative modeling leverages neural networks to approximate the score - the gradient of the log probability of the data - which enables efficient sampling in high dimensional space from learned distributions. Specifically, given a data distribution $p(\boldsymbol{x})$, a neural network is trained to approximate the score of the distribution $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$. Typically, the goal of score-based generative modeling is to generate samples from a data distribution $p(\boldsymbol{x})$ by solving the following reverse-diffusion stochastic differential equation (Anderson, 1982; Song et al., 2021),

$$d\boldsymbol{x} = \left(f(\boldsymbol{x}, t) - g(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right) dt + g(t) d\boldsymbol{w}, \quad (2)$$

where $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is the probability distribution of $\boldsymbol{x}$ at time t. To solve equation (2), a neural network $s(\boldsymbol{x}, t)$ is trained via denoising score matching (Vincent, 2011; Song et al., 2021; Hyvärinen, 2005) to approximate the correct score of the data distribution $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ for different levels of noise parameterized by a time variable t.

The data generation process achieved by solving equation (2) can be extended to sampling conditional distributions $p(\boldsymbol{x}|\boldsymbol{y})$ by training a neural network to approximate the conditional score $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}|\boldsymbol{y})$. In this case, the neural network is fed the conditional variables $\boldsymbol{y}$ as additional inputs.

In this work, we train a neural network, denoted as $s(\boldsymbol{x}, \boldsymbol{y}, t)$, to learn the score of the conditional distribution $p_t(\boldsymbol{x}|\boldsymbol{y})$ of galaxy count fields $\boldsymbol{x}$ obtained from computationally expensive hydro-simulations conditioned on computationally inexpensive N-body simulations $\boldsymbol{y}$. By learning this distribution, we can solve equation (2) to efficiently generate realizations of galaxy count fields, and circumvent the computational cost of running direct hydro-simulations.

For sampling, we choose to use the Variance-Exploding Stochastic Differential Equation, which is characterized by $f(\boldsymbol{x}, t) = 0$ and $g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$. The noise level during sampling is set by $\sigma(t) = \sigma_{\text{min}} \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t$. We explain how the parameters σ_{min} and σ_{max} are set in section 3.2.

3.1. Network Architecture

Following the work of (Legin et al., 2023) we adopt a U-net architecture (Ronneberger et al., 2015) with 3D convolutions to predict the conditional score distribution $p(\boldsymbol{x}|\boldsymbol{y})$. We closely follow the architecture code implemented in the package `score-models`[1], specifically, our model contains 4 downsampling/upsampling levels, each featuring 2 residual blocks derived from the `BigGAN` model (Brock et al., 2018). The base number of feature maps is 32, and is doubled on the third and fourth downsampling step. We train the network for 50,000 epochs, with a batch size of 20, on a single NVIDIA A100 40GB GPUs. The network weights are optimized with the Adam optimizer (Kingma & Ba, 2015) with a fixed learning rate of 2×10^{-4}. We apply gradient clipping to restrict the weight gradients to a maximum norm of 1. The entire duration of training is approximately 70 hours for 2.25 million iterations.

To incorporate the condition variable $\boldsymbol{y}$, we introduce it as an additional channel in the network's input. Since the values in N-body simulations are to the power of 10 we take a base-10 logarithmic in order to have values ranging from roughly 12 to 16 allowing for training to be more stable. Since we do not directly input the cosmological and astrophysical parameters to the neural network, it learns the score of the conditional galaxy count distribution marginalized over these parameters.

3.2. Diffusion and Discrete Data

Our goal is to learn the number of galaxies at each voxel which is an inherently discrete value. In recent years, frameworks for discrete diffusion have been proposed (Santos et al., 2023; Austin et al., 2023; Campbell et al., 2022); however their scalability and ability to perform as well as their continuous counterpart is still unclear. Therefore, we preferred developing our work with the well established continuous space denoising model.

However, we observed that directly applying continuous space models to discrete data led to inaccurate results specifically being unable to predict high valued galaxy counts in the hundreds. To mediate this issue we decided to log our galaxy counts on a base-10 which made the galaxy counts range within zero and three.

[1] github.com/AlexandreAdam/score_models

Based on this transformed data, we set the noising parameters σ_{max} to the largest euclidean distance in the training set as proposed by (Song & Ermon, 2020), which gave $\sigma_{\text{max}} = 44$. As for σ_{min} it is set to 1×10^{-5} which is below the integer precision needed to identify individual galaxy counts when reversing the log transform. Based on these two parameters, we set the number of discretized steps needed to sample between t∈[1,0] when solving equation (2) to 3000 in order to have a 96% change of staying in distribution (Song & Ermon, 2020).

Furthermore, we follow the idea of passing high frequency information to our network in order to better learn small scale information and better predict values of individual voxels (Kingma et al., 2021). To do so we pass our logged galaxy data through a sin and cos function which have a set frequency of 64 and 128. These frequencies were seen to give the best results in (Kingma et al., 2021), and these four additional maps are given as extra channels to the network. This procedure helped in better modeling sudden low to high value changes between nearby voxels.

4. Simulations

We used the Latin Hypercube (LH) set of simulations from the CAMELS-ASTRID suite (Villaescusa-Navarro et al., 2022). This dataset includes 1000 pairs of N-body and hydrodynamical simulations, each at a volume of 25 $(\text{Mpc}/h)^3$ and run using unique configuration of cosmological and astrophysical parameters. These parameters span a uniform prior over specified ranges: $\Omega_M \in [0.1, 0.5]$, $\sigma_8 \in [0.6, 1.0]$, $A_{\text{SN1}} \in [0.25, 4]$, $A_{\text{SN2}} \in [0.5, 2]$, $A_{\text{AGN1}} \in [0.25, 4]$, $A_{\text{AGN2}} \in [0.25, 4]$.

We identify galaxies in the hydro-simulation by using the provided catalogue of halos and subhalos which were found with the Friends of Friends (Davis et al., 1985) and SubFind (Springel et al., 2001) methods respectively. The chosen criteria for a subhalo to host a galaxy is containing stellar mass and having a mass greater than $10^{9.5}\ M_\odot/h$. We then place the galaxy catalogue on a 3D grid of size 32^3. Similarly, for the N-body simulation, the dark matter mass is set on a 3D grid of size 32^3. We divide the pair of gridded simulation into training and testing sets, allocating 900 simulations for training and 100 simulations for testing.

5. Results

For a given dark matter N-body simulation $\boldsymbol{y}$, we generate 100 samples of galaxy counts $\boldsymbol{x}$ from the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$. This is achieved by solving equation (2) using our trained score network $s(\boldsymbol{x}, \boldsymbol{y}, t)$. Our testing set contains 100 different N-body simulations leading to a total of 10,000 samples generated using 20 NVIDIA A100 40GB GPUs in parallel within 80 minutes.

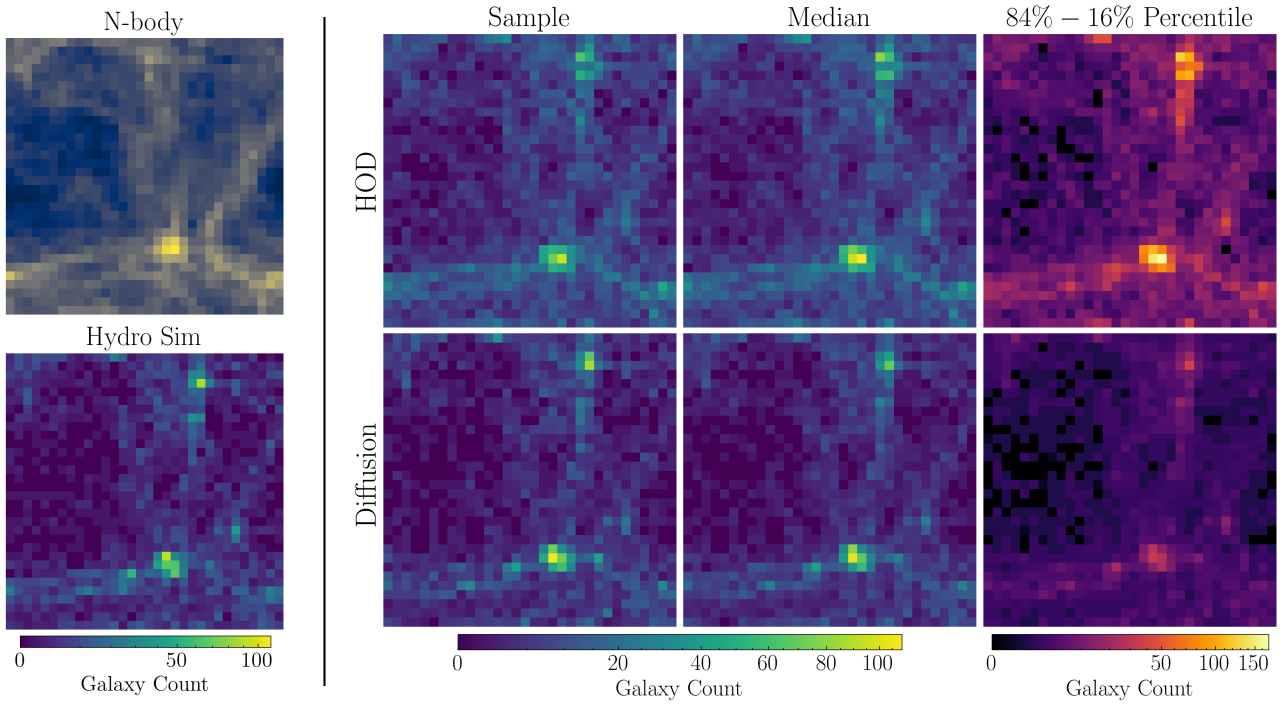

Figure 1. Comparison of galaxy count fields from a hydrodynamical simulation with predictions made by the Halo Occupation Distribution (HOD) method and our score-based generative diffusion model based on the corresponding dark matter N-body simulation. Each of the 2D fields showcased were obtained by summing the fields over the depth of the full 3D volume. The parameters used for the hydro-simulation are $\Omega_M = 0.25140$, $\sigma_8 = 0.8606$, $A_{\rm SN1} = 0.73002$, $A_{\rm AGN1} = 1.13131$, $A_{\rm SN2} = 0.70466$, $A_{\rm AGN2} = 1.07848$. The leftmost column shows the dark-matter-only N-body and corresponding hydro-simulation, ran at the same cosmology and initial conditions. We then show the median and 84th-16th percentile range of the galaxy counts for both the HOD and diffusion models. The N-body is shown for illustrative purposes and its color bar is not presented. The 84th-16th percentile range has a different color bar to better reflect the spread of galaxy counts predictions made by the HOD. This comparison highlights the differences in spatial distributions of galaxy counts between the two models and the improved prediction of the diffusion model over HOD.

In order to assess the validity of our samples we compare them to the hydro-simulation associated with the N-body simulation and the traditional HOD method. We take a Bayesian approach to best fit the parameters of the HOD to the power spectra of the fields over multiple cosmologies. Uniform priors are set on all 5 parameters of the HOD within the ranges $\log M_{\rm min} \in [9, 12]$, $\sigma_{\rm logM} \in [0.1, 0.6]$, $\log M_0 \in [11, 16]$, $\log M_1 \in [11, 16]$ and $\alpha \in [0, 1.5]$. An analytic likelihood for the power spectra at each cosmology cannot be used for the CAMELS-Astrid LH dataset as it only has one simulation per cosmology. Therefore, a Neural Likelihood Estimator (NLE) (Papamakarios et al., 2019) is trained on the power spectras of the 900 hydro-simulations in the training set. This allows the NLE to learn the conditional probability of HOD parameters which best reproduces the hydro-simulation power spectra over the different cosmologies in the training set. Finally, with the uniform priors and trained NLE, one hundred parameters from the posterior distribution are sampled using Markov Chain Monte Carlo at each cosmology in the testing set.

We first present the results for a single N-body simulation at a fixed cosmology in the test set. Figure 1 displays samples derived from the learned distribution alongside those generated by the HOD parametric method, accompanied by the corresponding true simulation of the galaxy count field. In Figure 2, we present a comparison of summary statistics of the generated galaxy count samples versus the ground truth galaxy count field from the hydrodynamical simulation showcased in Figure 1 (see additional examples of power spectra obtained for varying cosmological parameters in Appendix A). This comparison specifically includes the power spectrum, the transfer function (defined as the ratio of power spectra between the generated samples and the ground truth galaxy count field), and the cross-correlation of the two fields.

In order to assess the accuracy of the learned conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$ over all cosmologies in the testing set, we employ the Test of Accuracy with Random Points (TARP) coverage test from Lemos et al. (2023). This test determines whether the set of generated samples $\boldsymbol{x}_{\rm samples}$ accurately

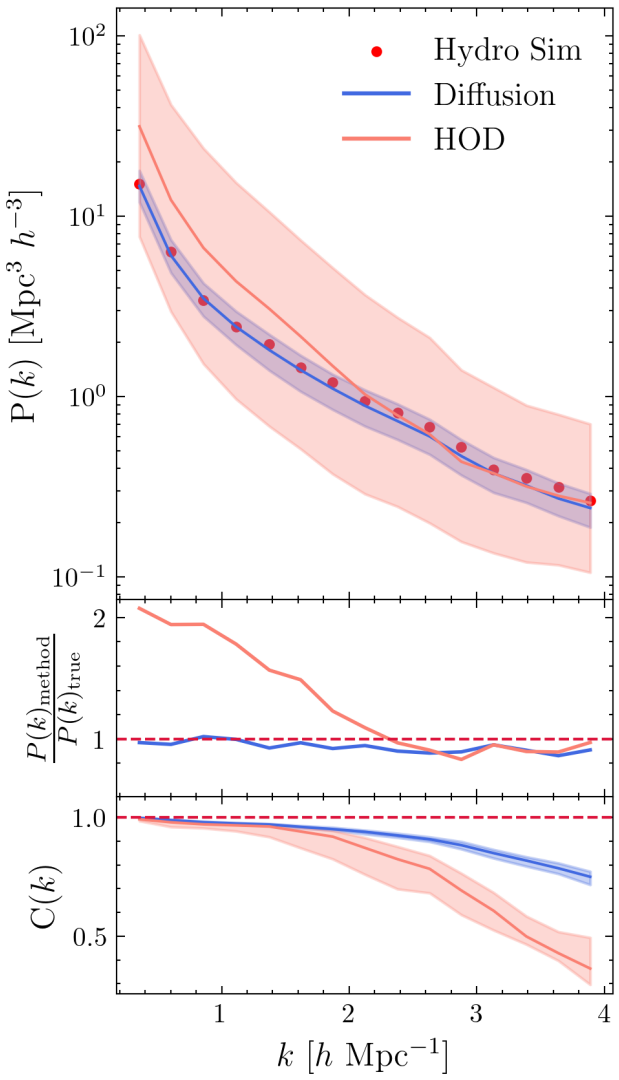

Figure 2. Summary statistics of the galaxy count fields for the results presented in Figure 1. The ground truth is the hydro-simulation $\boldsymbol{x}_{\text{true}}$ in red while conditional samples $p(\boldsymbol{x}|\boldsymbol{y})$ generated using the score network and the HOD are presented in blue and orange respectively. From top to bottom, for each method we plot the power spectra, power spectra ratio between the ground truth and generated samples, and the cross-correlation between the ground truth and generated samples. The median of the power and cross-correlation is presented for the diffusion and HOD while the uncertainty region is traced by the 16th and 84th percentile using 100 samples for both methods.

reflects the true underlying probability distribution $p(\boldsymbol{x}|\boldsymbol{y})$ by measuring the coverage probability of the samples over randomly-generated credible regions. We perform a TARP test on the power spectra of the fields with the results being depicted in Figure 3. The optimal case is a straight diagonal line indicating samples being generated within the underlying distribution formed by the hydro-simulation.

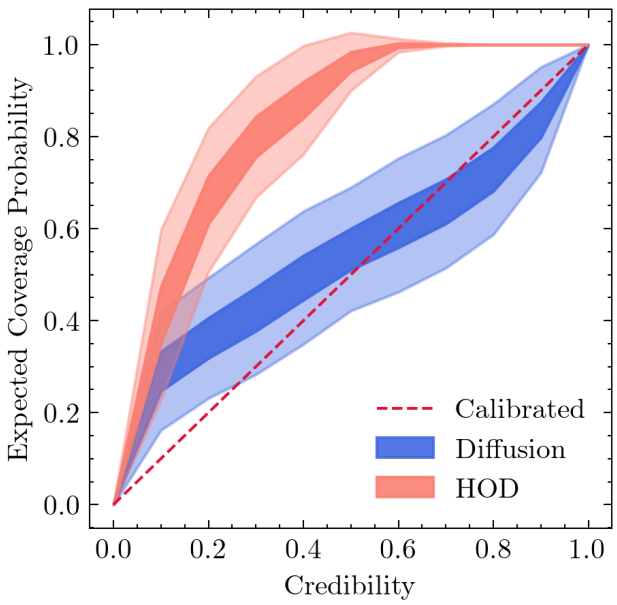

Figure 3. Coverage probability test using the TARP method to assess the accuracy of power spectra from samples generated by the diffusion model and the HOD method over the 100 hydro-simulations from the test set. The blue and orange curves represent the results for the diffusion model and HOD, respectively, with shaded areas indicating the 1σ and 3σ error obtained by bootstrapping examples from the test set. The dashed red line indicates the ideal calibrated line signifying accurate power spectra predictions. The diffusion model results show improved alignment with the diagonal compared to the HOD method, which exhibits a substantial deviation, indicating less accurate coverage.

6. Discussion

Our score-based generative model addresses the limitations of previous methods for predicting galaxy count fields from N-body simulations by learning the score $\nabla \log p_t(\boldsymbol{x}|\boldsymbol{y})$ of the conditional probability distribution of galaxy counts $\boldsymbol{x}$ and employing a reverse-diffusion process to generate samples from this distribution conditioned on dark matter N-body simulation $\boldsymbol{y}$. The model is benchmarked against the halo occupation distribution (HOD) model, which also employs probabilistic sampling but relies on significant assumptions about the distribution of galaxy counts.

A critical shortcoming of the HOD method is its failure to incorporate information about the local environment. In contrast, the U-net architecture used for learning the score of the distribution mitigates this by analyzing the field at multiple resolution levels, and is therefore capable of extracting information from a wider spatial region. Additionally, HOD-based predictions are not dependent on cosmological or astrophysical parameter values, whereas in principle, score-based generative models are capable of extracting and marginalizing over this information from dark matter N-

body simulations. This incorporation of local environmental factors and cosmological parameters renders score-based generative models as a more flexible approach for generating accurate galaxy count fields. A demonstration of this can be seen from Figure 1, where the fields generated by HOD appear to struggle to predict galaxy count features in regions of low density. Furthermore, it can be seen from Figure 1 that HOD predicts an overly high galaxy counts in certain regions of high density. In comparison, the samples generated with our diffusion model are capable of reliably replicating features which are consistent with the hydro-simulation in both low and high density regions. Additionally, our model is able to accurately capture sudden changes in density as can be seen with the voids in Figure 1.

The limitations of the HOD model are further illustrated by our results in Figure 2, which presents various summary statistics of the fields. As shown, the score-based generative model produces samples with power spectra that accurately match the true hydro-simulation across all relevant scales. In contrast, the HOD model is under confident across all scales with a bias at lower k modes. The cross-correlation displayed in Figure 2 also highlights our model's ability to interpolate positional differences between high density regions in N-body and hydro-simulations. The cross-correlation is sensitive to the position of features in images as it captures their Fourier phase. It is known that halo positions slightly differ in N-body simulations compared to hydro-simulations due to the gravitational effects of the additional baryons. The HOD cannot correct for this difference in halo positions, however, the fact that our model's cross-correlation is higher than the HOD for all scales highlights its ability to correct this discrepancy.

Although the ideal cross-correlation is one across all scales, it is important to note that the decrease in cross-correlation at higher k modes for both the diffusion model samples and the HOD samples is expected, as non-linear processes dominate making it inherently more difficult in predicting the exact spatial distribution of galaxies solely from dark matter density fields. Regardless, in future work, we are interested in examining whether the score-generative model is able to extract the full information content available in dark matter N-body density fields for predicting the spatial distribution of galaxies.

While Figure 2 visually assesses the accuracy of the HOD model compared to the diffusion model for a single hydrodynamical simulation, Figure 3 quantitatively evaluates the accuracy of the power spectra across all 100 hydrodynamical simulation examples in the test set. Specifically, Figure 3 illustrates whether the power spectra of samples generated from the HOD and diffusion models accurately represent the distribution of possible power spectra of hydrodynamical galaxy count fields. Our diffusion model, which is closer to the diagonal line indicating perfect accuracy, demonstrates a significant improvement over the HOD model. However, our model still exhibits a bias at lower credibility values, indicating the need for further training improvements. This bias may result from the insufficient number of training examples (900 hydro-simulations) used in training the diffusion model. To address this, future work will explore data augmentation and pretraining on additional cosmological simulation datasets to enhance the model's generalization capabilities.

The primary objective of this work is to evaluate the potential of score-based generative models as reliable emulators of hydrodynamical fields, such as galaxy counts, at a fraction of the computational cost. Achieving this would facilitate cosmological inference tasks where many hydrodynamical simulations need to be generated to compare different models with observed galaxy distributions. By directly conditioning on the dark matter density field, this method also represents a significant step towards making forward models differentiable in cosmological simulations, completely removing the need to run a non-differentiable halo finder algorithm (Davis et al., 1985; Springel et al., 2001). Consequently, to assess the viability of our diffusion model for cosmological inference, we aim to determine whether galaxy count samples generated by our diffusion model are sufficiently accurate to serve as training data for simulation-based inference methods to predict unbiased cosmological parameter values. Specifically, we will verify if the predictions derived from our model's samples are consistent with those obtained from real hydrodynamical simulations. This validation would demonstrate the potential of score-based generative models as an efficient method for producing training data of hydrodynamical simulations that are otherwise computationally prohibitive.

Acknowledgements

This project was developed as part of the Simons Collaboration on "Learning the Universe." The Flatiron Institute is supported by the Simons Foundation. The work is in part supported by computational resources provided by Calcul Quebec and the Digital Research Alliance of Canada. A.B. and M.H. are supported by the Simons Collaboration on "Learning the Universe." Y. H. and L. P.-L. acknowledge support from the Canada Research Chairs Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) through grants RGPIN- 2020-05073 and 05102, and the Fonds de recherche du Québec through grants 2022-NC-301305 and 300397. R.L. and A.A. are supported by NSERC CGS D scholarship.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces, 2023.

Baldi, M., Pettorino, V., Robbers, G., and Springel, V. Hydrodynamical n-body simulations of coupled dark energy cosmologies. *Monthly Notices of the Royal Astronomical Society*, 403(4):1684–1702, 2010.

Bird, S., Ni, Y., Di Matteo, T., Croft, R., Feng, Y., and Chen, N. The ASTRID simulation: galaxy formation and reionization. *Monthly Notices of the Royal Astronomical Society*, 512(3):3703–3716, 03 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac648. URL https://doi.org/10.1093/mnras/stac648.

Bozorgnia, N. and Bertone, G. Implications of hydrodynamical simulations for the interpretation of direct dark matter searches. *International Journal of Modern Physics A*, 32(21):1730016, 2017.

Brock, A., Donahue, J., and Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv e-prints*, art. arXiv:1809.11096, September 2018. doi: 10.48550/arXiv.1809.11096.

Campbell, A., Benton, J., Bortoli, V. D., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models, 2022.

Chen, N., Ni, Y., Tremmel, M., Di Matteo, T., Bird, S., DeGraf, C., and Feng, Y. Dynamical friction modelling of massive black holes in cosmological simulations and effects on merger rate predictions. *Monthly Notices of the Royal Astronomical Society*, 510(1):531–550, 2022.

Cuesta-Lazaro, C. and Mishra-Sharma, S. A point cloud approach to generative modeling for galaxy surveys at the field level. *arXiv preprint arXiv:2311.17141*, 2023.

Davis, M., Efstathiou, G., Frenk, C. S., and White, S. D. The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 292, May 15, 1985, p. 371-394. Research supported by the Science and Engineering Research Council of England and NASA.*, 292:371–394, 1985.

Davé, R., Anglés-Alcázar, D., Narayanan, D., Li, Q., Rafieferantsoa, M. H., and Appleby, S. simba: Cosmological simulations with black hole growth and feedback. *Monthly Notices of the Royal Astronomical Society*, 486(2):2827–2849, 04 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz937. URL https://doi.org/10.1093/mnras/stz937.

de Santi, N. S., Shao, H., Villaescusa-Navarro, F., Abramo, L. R., Teyssier, R., Villanueva-Domingo, P., Ni, Y., Anglés-Alcázar, D., Genel, S., Hernández-Martínez, E., et al. Robust field-level likelihood-free inference with galaxies. *The Astrophysical Journal*, 952(1):69, 2023.

Desjacques, V., Jeong, D., and Schmidt, F. Large-scale galaxy bias. *Physics reports*, 733:1–193, 2018.

Garcia-Quintero, C., Mena-Fernández, J., Rocher, A., Yuan, S., Hadzhiyska, B., Alves, O., Rashkovetskyi, M., Seo, H., Padmanabhan, N., Nadathur, S., et al. Hod-dependent systematics in emission line galaxies for the desi 2024 bao analysis. *arXiv preprint arXiv:2404.03009*, 2024.

Hadzhiyska, B., Bose, S., Eisenstein, D., Hernquist, L., and Spergel, D. N. Limitations to the 'basic' HOD model and beyond. *Monthly Notices of the Royal Astronomical Society*, 493(4):5506–5519, 03 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa623. URL https://doi.org/10.1093/mnras/staa623.

Heitmann, K., Finkel, H., Pope, A., Morozov, V., Frontiere, N., Habib, S., Rangel, E., Uram, T., Korytov, D., Child, H., et al. The outer rim simulation: A path to many-core supercomputers. *The Astrophysical Journal Supplement Series*, 245(1):16, 2019.

Ho, M., Soltis, J., Farahi, A., Nagai, D., Evrard, A., and Ntampaka, M. Benchmarks and explanations for deep learning estimates of x-ray galaxy cluster masses. *Monthly Notices of the Royal Astronomical Society*, 524 (3):3289–3302, 2023.

Ho, M., Bartlett, D. J., Chartier, N., Cuesta-Lazaro, C., Ding, S., Lapel, A., Lemos, P., Lovell, C. C., Makinen, T. L., Modi, C., et al. Ltu-ili: An all-in-one framework for implicit inference in astrophysics and cosmology. *arXiv preprint arXiv:2402.05137*, 2024.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b578f2a52a0229873fefc2a4b06377fa-Paper.pdf.

Kugel, R., Schaye, J., Schaller, M., Helly, J. C., Braspenning, J., Elbers, W., Frenk, C. S., McCarthy, I. G., Kwan, J., Salcido, J., et al. Flamingo: calibrating large cosmological hydrodynamical simulations with machine learning. *Monthly Notices of the Royal Astronomical Society*, 526 (4):6103–6127, 2023.

Legin, R., Ho, M., Lemos, P., Perreault-Levasseur, L., Ho, S., Hezaveh, Y., and Wandelt, B. Posterior sampling of the initial conditions of the universe from non-linear large scale structures using score-based generative models. *Monthly Notices of the Royal Astronomical Society: Letters*, 527(1):L173–L178, 10 2023. ISSN 1745-3925. doi: 10.1093/mnrasl/slad152. URL https://doi.org/10.1093/mnrasl/slad152.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-based accuracy testing of posterior estimators for general inference. In *International Conference on Machine Learning*, pp. 19256–19273. PMLR, 2023.

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., Hadzhiyska, B., Bose, S., and Satterthwaite, T. P. Abacussummit: a massive set of high-accuracy, high-resolution n-body simulations. *Monthly Notices of the Royal Astronomical Society*, 508(3):4017–4037, 2021.

National Academies of Sciences, Engineering, and Medicine. *Pathways to Discovery in Astronomy and Astrophysics for the 2020s*. The National Academies Press, Washington, DC, 2023. ISBN 978-0-309-46734-6. doi: 10.17226/26141.

Nelson, D., Springel, V., Pillepich, A., Rodriguez-Gomez, V., Torrey, P., Genel, S., Vogelsberger, M., Pakmor, R., Marinacci, F., Weinberger, R., Kelley, L., Lovell, M., Diemer, B., and Hernquist, L. The illustristng simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1):2, 05 2019. ISSN 2197-7909. doi: 10.1186/s40668-019-0028-x. URL https://doi.org/10.1186/s40668-019-0028-x.

Ono, V., Park, C. F., Mudur, N., Ni, Y., Cuesta-Lazaro, C., and Villaescusa-Navarro, F. Debiasing with diffusion: Probabilistic reconstruction of dark matter fields from galaxies with camels. *arXiv preprint arXiv:2403.10648*, 2024.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 837–848. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/papamakarios19a.html.

Richardson, J., Zheng, Z., Chatterjee, S., Nagai, D., and Shen, Y. The halo occupation distribution of sdss quasars. *The Astrophysical Journal*, 755(1):30, 2012.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1505.html#RonnebergerFB15.

Rouhiainen, A., Gira, M., Münchmeyer, M., Lee, K., and Shiu, G. Super-resolution emulation of large cosmological fields with a 3d conditional diffusion model. *arXiv preprint arXiv:2311.05217*, 2023.

Santos, J. E., Fox, Z. R., Lubbers, N., and Lin, Y. T. Blackout diffusion: Generative diffusion models in discrete-state spaces, 2023.

Schaye, J., Kugel, R., Schaller, M., Helly, J. C., Braspenning, J., Elbers, W., McCarthy, I. G., van Daalen, M. P., Vandenbroucke, B., Frenk, C. S., et al. The flamingo project: cosmological hydrodynamical simulations for large-scale structure and galaxy cluster surveys. *Monthly Notices of the Royal Astronomical Society*, 526(4):4978–5020, 2023.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Springel, V., White, S. D., Tormen, G., and Kauffmann, G. Populating a cluster of galaxies–i. results at z= 0. *Monthly Notices of the Royal Astronomical Society*, 328 (3):726–750, 2001.

Villaescusa-Navarro, F., Hahn, C., Massara, E., Banerjee, A., Delgado, A. M., Ramanah, D. K., Charnock, T., Giusarma, E., Li, Y., Allys, E., et al. The quijote simulations. *The Astrophysical Journal Supplement Series*, 250 (1):2, 2020.

Villaescusa-Navarro, F., Genel, S., Anglé s-Alcázar, D., Thiele, L., Dave, R., Narayanan, D., Nicola, A., Li, Y., Villanueva-Domingo, P., Wandelt, B., Spergel, D. N., Somerville, R. S., Matilla, J. M. Z., Mohammad, F. G., Hassan, S., Shao, H., Wadekar, D., Eickenberg, M., Wong, K. W. K., Contardo, G., Jo, Y., Moser, E., Lau, E. T., Valle, L. F. M. P., Perez, L. A., Nagai, D., Battaglia, N., and Vogelsberger, M. The CAMELS multifield data set: Learning the universe's fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2):61, 04 2022. doi: 10.3847/1538-4365/ac5ab0. URL https://doi.org/10.3847%2F1538-4365%2Fac5ab0.

Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., Perez, L. A., Villanueva-Domingo, P., Wadekar, D., Shao, H., Mohammad, F. G., Hassan, S., Moser, E., et al. The camels project: public data release. *The Astrophysical Journal Supplement Series*, 265(2):54, 2023.

Vincent, P. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7): 1661–1674, 07 2011. ISSN 0899-7667. doi: 10.1162/NECO_a_00142. URL https://doi.org/10.1162/NECO_a_00142.

Vogelsberger, M., Genel, S., Springel, V., Torrey, P., Sijacki, D., Xu, D., Snyder, G., Nelson, D., and Hernquist, L. Introducing the illustris project: simulating the coevolution of dark and visible matter in the universe. *Monthly Notices of the Royal Astronomical Society*, 444(2):1518–1547, 2014.

Vogelsberger, M., Marinacci, F., Torrey, P., and Puchwein, E. Cosmological simulations of galaxy formation. *Nature Reviews Physics*, 2(1):42–66, 2020.

Yuan, S., Garrison, L. H., Hadzhiyska, B., Bose, S., and Eisenstein, D. J. Abacushod: a highly efficient extended multitracer hod framework and its application to boss and eboss data. *Monthly Notices of the Royal Astronomical Society*, 510(3):3301–3320, 2022.

Zhang, X., Wang, Y., Zhang, W., Sun, Y., He, S., Contardo, G., Villaescusa-Navarro, F., and Ho, S. From dark matter to galaxies with convolutional networks, 2019.

Zheng, Z., Berlind, A. A., Weinberg, D. H., Benson, A. J., Baugh, C. M., Cole, S., Davé, R., Frenk, C. S., Katz, N., and Lacey, C. G. Theoretical models of the halo occupation distribution: Separating central and satellite galaxies. *The Astrophysical Journal*, 633(2):791, 11 2005. doi: 10.1086/466510. URL https://dx.doi.org/10.1086/466510.

Zheng, Z., Coil, A. L., and Zehavi, I. Galaxy evolution from halo occupation distribution modeling of deep2 and sdss galaxy clustering. *The Astrophysical Journal*, 667 (2):760, oct 2007. doi: 10.1086/521074. URL https://dx.doi.org/10.1086/521074.

A. Appendix

We present further summary statistics at different cosmologies to demonstrate our model's ability to generalize.

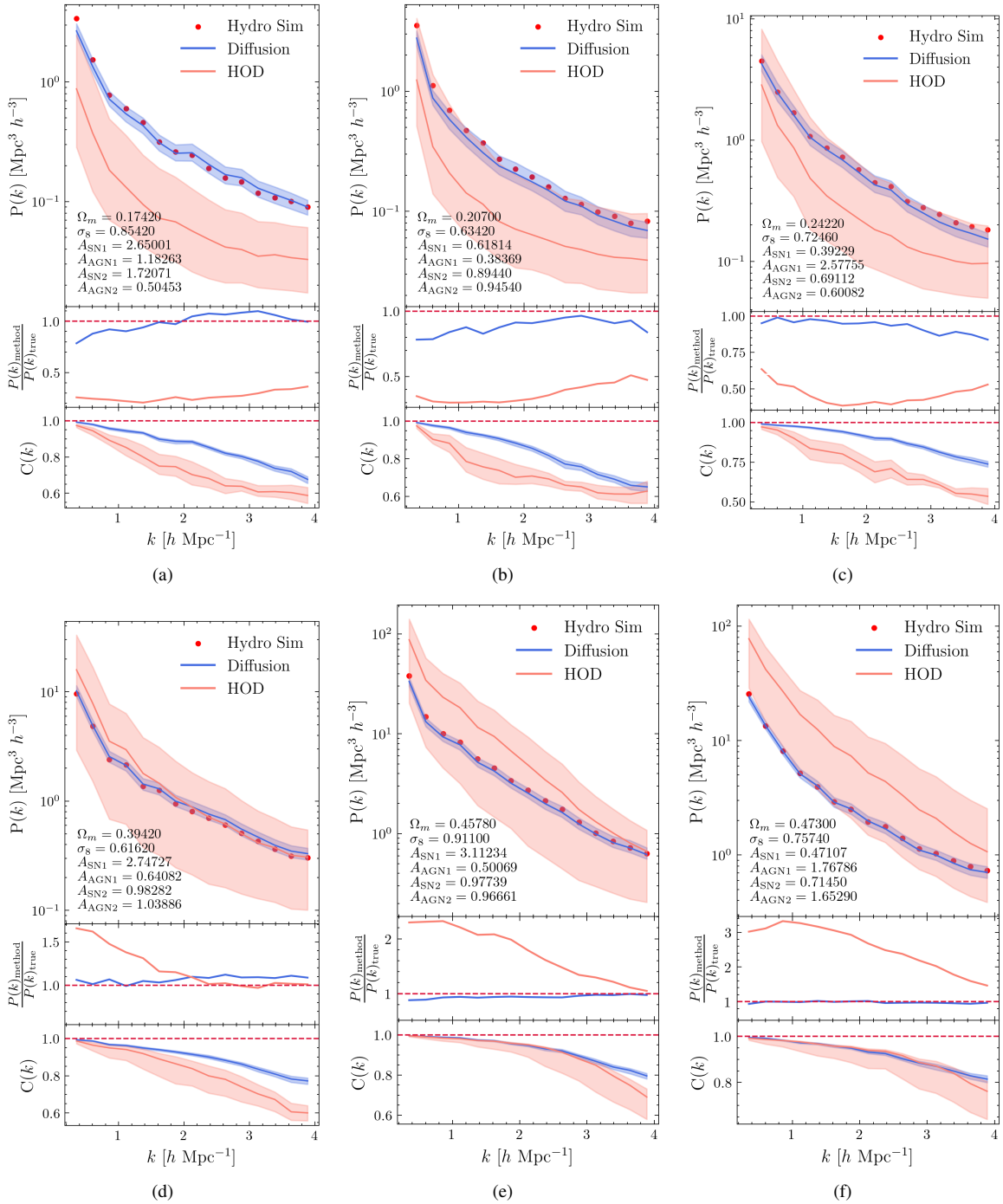

Figure 4. Summary statistics from various hydrodynamical simulations from the test set, encompassing a broad range of cosmological parameter values, are presented alongside results from samples generated using both the HOD method and the diffusion model. These results demonstrate that the diffusion model effectively generalizes across a wide range of cosmological parameter values beyond the training set. In contrast, the HOD method shows inferior generalization capabilities for simulations with different cosmological parameter values showcased here.