Field-level simulation-based inference of galaxy clustering with convolutional neural networks

Pablo Lemos,[1,2,3,4,*] Liam Parker,[5,*] ChangHoon Hahn,[5] Shirley Ho,[4,5,6,7] Michael Eickenberg,[8] Jiamin Hou,[9,10] Elena Massara,[11,12] Chirag Modi,[4,8] Azadeh Moradinezhad Dizgah,[13] Bruno Régaldo-Saint Blancard,[8] and David Spergel[4,5]

(SimBIG Collaboration)

[1]*Department of Physics, Université de Montréal, Montréal, 1375 Avenue Thérèse-Lavoie-Roux, Montréal, QC H2V 0B3, Canada*
[2]*Mila—Quebec Artificial Intelligence Institute, Montréal, 6666 Rue Saint-Urbain, Montréal, QC H2S 3H1, Canada*
[3]*Ciela—Montreal Institute for Astrophysical Data Analysis and Machine Learning, Montréal, Canada*
[4]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, New York 10010, USA*
[5]*Department of Physics, Princeton University, Princeton, New Jersey 08544, USA*
[6]*Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, New York 10003, USA*
[7]*Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*
[8]*Center for Computational Mathematics, Flatiron Institute, 162 5th Avenue, New York, New York 10010, USA*
[9]*Department of Astronomy, University of Florida, 211 Bryant Space Science Center, Gainesville, Florida 32611, USA*
[10]*Max-Planck-Institut für Extraterrestrische Physik, Postfach 1312, Giessenbachstrasse 1, 85748 Garching bei München, Germany*
[11]*Waterloo Centre for Astrophysics, University of Waterloo, 200 University Avenue W., Waterloo, Ontario N2L 3G1, Canada*
[12]*Department of Physics and Astronomy, University of Waterloo, 200 University Avenue W., Waterloo, Ontario N2L 3G1, Canada*
[13]*Département de Physique Théorique, Université de Genève, 24 quai Ernest Ansermet, 1211 Genève 4, Switzerland*

(Received 13 July 2023; accepted 11 December 2023; published 30 April 2024)

We present the first simulation-based inference (SBI) of cosmological parameters from field-level analysis of galaxy clustering. Standard galaxy clustering analyses rely on analyzing summary statistics, such as the power spectrum P_ℓ, with analytic models based on perturbation theory. Consequently, they do not fully exploit the nonlinear and non-Gaussian features of the galaxy distribution. To address these limitations, we use the SimBIG forward modeling framework to perform SBI using normalizing flows. We apply SimBIG to a subset of the Baryon Oscillation Spectroscopic Survey CMASS galaxy sample using a convolutional neural network with stochastic weight averaging to perform massive data compression of the galaxy field. We infer constraints on $\Omega_m = 0.267^{+0.033}_{-0.029}$ and $\sigma_8 = 0.762^{+0.036}_{-0.035}$. While our constraints on Ω_m are in line with standard P_ℓ analyses, ours on σ_8 are 2.65× tighter. Our analysis also provides constraints on the Hubble constant $H_0 = 64.5 \pm 3.8$ km/s/Mpc from galaxy clustering alone. This higher constraining power comes from additional non-Gaussian cosmological information, inaccessible with P_ℓ. We demonstrate the robustness of our analysis by showcasing our ability to infer unbiased cosmological constraints from a series of test simulations that are constructed using different forward models than the one used in our training dataset. This work not only presents competitive cosmological constraints but also introduces novel methods for leveraging additional cosmological information in upcoming galaxy surveys like the Dark Energy Spectroscopic Instrument, Prime Focus Spectrograph, and *Euclid*.

DOI: 10.1103/PhysRevD.109.083536

[*]These authors contributed equally to this work.

I. INTRODUCTION

Precision measurements of cosmological parameters, such as the matter density and the expansion rate of the Universe, play a crucial role in shaping our understanding of the evolution and structure of the cosmos. These parameters can be inferred from a variety of observational data, including measurements of the statistical properties of the large-scale structure of the Universe traced by the distribution of galaxies.

Traditionally, cosmological parameter inference has relied on analyzing the distribution of galaxies using summary statistics—most often the power spectrum, $P_\ell(k)$ (e.g., [1–9]). In addition, these analyses incorporate analytical modeling of galaxy clustering through perturbation theory (PT; see [10,11] for reviews). Consequently, these analyses have been limited to large, weakly nonlinear scales where the deviation from PT is small. By only considering the power spectrum, these analyses cannot exploit the rich non-Gaussian information in the galaxy distribution, which is only weakly imprinted on the power spectrum.

Recent analyses of BOSS data have now established that there is in fact significant non-Gaussian cosmological information on nonlinear scales in galaxy clustering. Furthermore, previous galaxy clustering analyses using higher-order clustering statistics have produced significantly tighter constraints than with P_ℓ alone (e.g., [12–15]). Furthermore, forecasts that employ various summary statistics beyond P_ℓ (e.g., [16–21]) have been shown to produce even tighter constraints by including nonlinear scales. Nonetheless, these applications remain limited by the inability of PT to model galaxy clustering at scales beyond the quasilinear, especially for higher-order statistics.

Another major challenge of galaxy clustering analyses is their inability to fully account for observational systematics. For example, fiber collisions have been shown to significantly bias P_ℓ on scales smaller than $k \sim 0.1\ h/\mathrm{Mpc}$ [22,23]. Observational effects in targeting, imaging, and completeness also significantly impact clustering measurements [24,25]. Finally, these analyses assume a Gaussian functional form of the likelihood function used in their Bayesian framework. This assumption does not necessarily hold in general [26–28].

To overcome these limitations, we instead use simulation-based inference[1] (SBI). SBI uses forward models of the observables, instead of analytic models, and then infers a posterior distribution over the parameters (or a likelihood, that can then be converted into the posterior with the Bayes theorem). This method enables us to leverage high-fidelity simulations that accurately model complex physical processes, leading to more robust inferences than methods based on analytical models.

There have already been multiple applications of SBI in astronomy (e.g., [29–44]). In the specific context of galaxy clustering, Hahn *et al.* [45] introduced the Simulation-Based Inference of Galaxies (SimBIG) forward models, which produce realistic mock observations of the Sloan Digital Sky Survey III Baryon Oscillation Spectroscopic Survey (BOSS [46,47]) Southern Galactic Cap (SGC) at different cosmologies and includes systematic effects such as survey geometry and fiber collisions. Using these models, we were able to robustly infer Λ cold dark matter (ΛCDM) parameters from the BOSS CMASS-SGC sample at scales down to $k_{\mathrm{max}} \sim 0.5\ h/\mathrm{Mpc}$. These works, however, focused on presenting the SimBIG framework and relied on compressing the galaxy distribution to the power spectrum, which does not capture the non-Gaussian information present.

In this work, we extend SimBIG to analyze the galaxy distribution directly at the field-level.[2] Specifically, we use convolutional neural networks (CNNs) to perform massive data compression and to extract maximally relevant features from the galaxy distribution. By learning the maximally relevant features with CNNs, our approach aims to extract even more cosmological information than summary statistics and establish a comprehensive framework for extracting *all* of the cosmological information in galaxy distributions.

The rest of the paper is organized as follows. We describe the details of the observations and simulations in Sec. II. Our methodology is explained in Sec. III and applied to observations in Sec. IV. Finally, we present our conclusions in Sec. V.

II. OBSERVATIONS AND SIMULATIONS

In this section, we describe the observational galaxy sample as well as the forward-modeled training and test simulations.

A. Observations: BOSS CMASS SGC

We use a sample of CMASS luminous red galaxies from the BOSS data release 12 as our observational data [48]. We limit our analysis to the subsample of CMASS sample at the Southern Galactic Cap within the angular footprint $\mathrm{DEC} > -6°$ and $-25° < \mathrm{RA} < 28°$ and redshift range $0.45 < z < 0.6$. The reason for this is that the Quijote simulation boxes are not big enough to include the full CMASS sample. In total, our sample consists of 109,636 galaxies. Visual illustrations of the sample can be found in [45,49].

B. SimBIG forward model

We use the SimBIG forward modeling pipeline to generate field-level synthetic observations that aim to be statistically

[1]The terms "likelihood-free inference" and "implicit likelihood inference" have also been used to refer to the same method.

[2]The term "field-level" is used here to refer to the fact that our neural network takes as input the field, as opposed to performing some compression step on the field before feeding it to the neural network, such as the power spectrum or bispectrum.

indistinguishable from BOSS observations. This pipeline consists of four distinct steps: (1) N-body simulations, (2) a dark matter halo finder, (3) a halo occupation distribution framework (HOD), and (4) application of survey realism.

The N-body simulations are taken from the Quijote suite [50]. The simulations evolve 1024^3 cold dark matter particles from $z = 127$ to $z = 0.5$ in a cosmological volume $1\ (h^{-1}\,\mathrm{Gpc})^3$ using the TreePM GADGET-III code. These simulations accurately model matter clustering down to nonlinear scales beyond $k = 0.5h/\mathrm{Mpc}$.

From these N-body simulations, dark matter halos are identified using the ROCKSTAR halo finder [51], which has been shown to robustly and accurately track dark matter halo location and substructure using phase-space information. Specifically, the standard [52] HOD model, which populates halos using M_h and five free HOD parameters, is expanded by including assembly, concentration, and velocity biases. These biases add the necessary flexibility to account for recent evidence suggesting that galaxies occupy halos in ways that depend on halo properties beyond M_h (e.g., assembly history; [53–56]).

Finally, survey realism is applied to the HOD galaxy catalog to produce a CMASS-like galaxy catalog. First, the $1(h^{-1}\mathrm{Gpc})^3$ box is remapped to a cuboid [57] and then cut to the BOSS survey geometry. Then, the galaxy catalog is trimmed to $z \in (0.45, 0.6)$, and fiber collisions are applied. Ultimately, the forward models are determined by five ΛCDM cosmological parameters, Ω_m, Ω_b, h, n_s, σ_8 and nine HOD parameters. We refer readers to Hahn *et al.* [45,49] for further details.

To construct our training set, we use 2518 high-resolution Quijote N-body simulations[3] arranged in a Latin hypercube configuration (LHC), which imposes priors on the cosmological parameters that conservatively encompass the *Planck* cosmological constraints. For each simulation, we forward-model ten CMASS-like galaxy catalogs using unique HOD parameters randomly sampled from a conservative prior. While this is suboptimal, as it leads to samples that are not independent and identically distributed (i.i.d.), this factor 10 increase in the number of available simulations greatly improves our results, and we expect regularization to deal with any potential issues arising from not i.i.d. samples. We split the resulting 25,180 simulations into a 20,000 and 5180 training and validation set.

C. Test simulations

In order to demonstrate that we can infer accurate and unbiased cosmological constraints, we test our analysis on three different sets of realistic test simulations that differ from the training dataset and have been developed within SimBIG and introduced in [49]: TEST0, TEST1, and TEST2.

TEST0 uses Quijote N-body simulations that have the same specifications as those arranged in the LHC, but were run at a fiducial cosmology with $\Omega_m = 0.3175$, $\Omega_b = 0.049$, $h = 0.6711$, $n_s = 0.9624$, $\sigma_8 = 0.834$. The halo finder, HOD framework, and survey realism are the same as those used in the training set, but the HOD parameters span a narrower prior. This test dataset contains 500 synthetic galaxy catalogs.

TEST1 involves the same N-body simulations as TEST0, but a different halo finder: the friend-of-friend (FoF) algorithm [58]. Assembly, concentration, and satellite velocity biases are also not considered in the HOD model. Central velocity bias is implemented, as the halo velocities in FoF halo catalogs correspond to the bulk velocity of the dark matter particles in the halo rather than the velocity of the central density peak of the halo. This test dataset contains 500 synthetic galaxy catalogs.

TEST2 uses 25 AbacusSummit N-body simulations [59] in the "base" configuration of the suite. The simulations contain 6912^3 particles in a $(2h^{-1}\mathrm{Gpc})^3$ volume box. Halo catalogs are constructed from these simulations using the CompaSO halo finder [60] and each of them is divided into eight boxes of volume $1(h^{-1}\mathrm{Gpc})^3$. Halos are populated with galaxies using the same HOD model implemented in the training set, with HOD parameters that sample the same narrower priors used in TEST0. This test dataset contains 1000 synthetic galaxy catalogs.

All three test datasets incorporate the same survey realism as the training dataset to produce CMASS-like galaxy catalogs.

It would be ideal to have a third set that we only tested on after passing validation tests on TEST0, TEST1, and TEST2. However, due to the high computational cost of our simulations, this was unfeasible.

D. Galaxy density field

To apply CNNs to our observational and simulated galaxy samples, we mesh the galaxy distribution into a box, with voxel size $64 \times 128 \times 128$. We choose this size because divisibility by 2 allows for easier downsampling in the CNN. First, we place the distribution into a $[707, 1414, 1414]\ \mathrm{Mpc}/h$ box and convert it into a 3D density field using a cloud-in-cell mass assignment [61]. For our observational sample, we include systematics weights for multiple effects (redshift failures, stellar density, and seeing conditions; [24,62]) in the mass assignment.

Since our data occupy a $[577.3, 1414, 1224]\ \mathrm{Mpc}/h$ box, we fill some of the box with zero-valued voxels. Our voxels have size $\sim[11, 11, 11]\ \mathrm{Mpc}/h$, thus we impose an effective scale cut of $k < k_{\mathrm{max}} = 0.28\ h/\mathrm{Mpc}$. While this is larger than the scale cut imposed in the SimBIG P_ℓ analysis [63], we find that it is sufficient to place significant cosmological constraints. Moreover, pushing to even smaller scale cuts presents its own set of challenges. For one, smaller scale cuts present significant computational challenges in terms

[3]We supplement the 2000 Quijote N-body simulations used in [45] with 518 additionally constructed simulations.

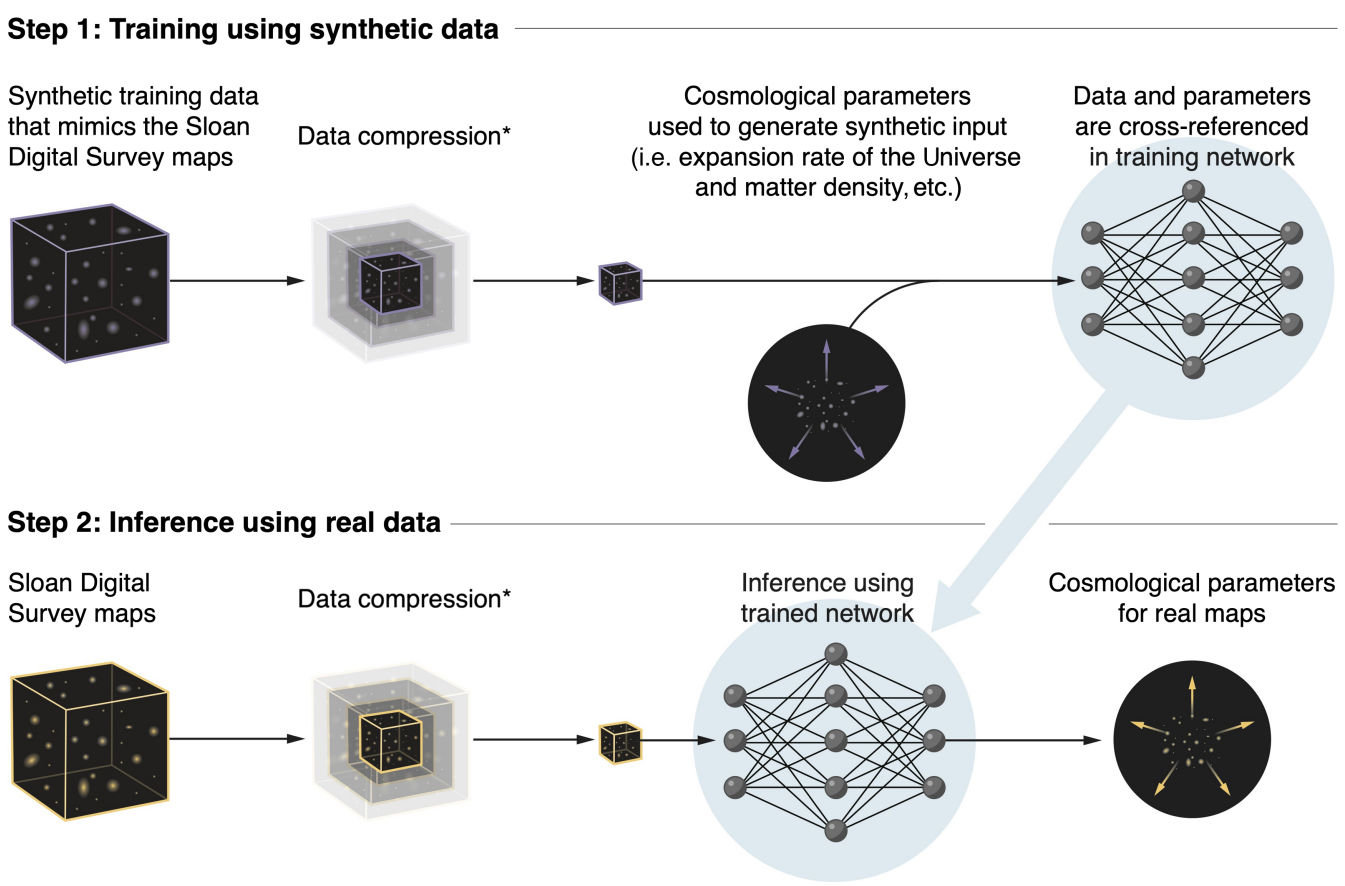


FIG. 1. Schematic illustrating the various elements of the SimBIG forward modeling pipeline. First, we generate synthetic galaxy catalogs that mimic the real BOSS observations. Then, we train a data compression step using our CNN to compress the catalog to its cosmological parameters. Next, we train a neural posterior estimator on the estimating parameters and the true parameters to estimate posteriors over the cosmological parameters. Once our data compression and neural posterior estimator are trained, we apply our pipeline to infer cosmological parameters from the real BOSS observations.

of the required memory to train on larger forward model sizes. Additionally, we find that models trained on smaller scale cuts tend to overfit on the training dataset significantly, limiting the robustness of their inferred parameters.

III. METHODS

Our approach to field-level inference of cosmological parameters consists of two main components: a massive data compression/feature extraction step performed by a CNN, followed by SBI. In the following section, we describe each step in more detail and also provide a visual description in Fig. 1.[4] We also describe two additional elements of our analysis, designed to ensure accurate posterior estimates: weight marginalization and validation with coverage probability tests.

A. CNN-based feature extraction

CNNs are flexible machine learning models that can be optimized to extract maximally relevant features from their inputs across a wide variety tasks. They consist of multiple layers of specialized kernels that are convolved across the input to extract features in a hierarchical scheme. These networks are particularly well suited for image-based tasks due to their ability to (1) exploit local receptive fields, (2) recognize patterns regardless of their position in the input due to translational invariance, and (3) extract increasingly complex features by combining lower-level features from previous layers hierarchically (for a review of CNNs, see [64]).

In this study, we train a three-dimensional CNN to compress the galaxy density fields produced by the SimBIG forward models to the cosmological parameters of those models. Specifically, the CNN takes as input the three-dimensional tensor representing the discretized forward model $x \in \mathbb{R}^{64\times128\times128}$ and outputs a prediction $\hat{\boldsymbol{\theta}}$ of the ΛCDM cosmological parameters, $\{\Omega_m, \Omega_b, h, n_s, \sigma_8\}$, used to generate that forward model.

[4]We also attempted a one-step approach, where the CNN served as embedding to the SBI step; however, we found that constraints were significantly weaker when using this approach.

The CNN architecture consists of five convolutional blocks. Each convolutional block begins with a convolutional layer that convolves its input with a number of $3 \times 3 \times 3$ kernels. This convolution is performed with 1-voxel zero padding. This is followed by a rectified linear unit (ReLU). The output of the ReLU unit is then downsampled using max pooling, which enables the network to learn features at increasing scales by reducing the size of its internal representations. Finally, batch normalization is applied, which typically speeds up training and has been shown to help with generalization [65]. Following the convolutional blocks, the activation maps are flattened and fed into three fully connected layers that output $\hat{\boldsymbol{\theta}}$. These layers also use ReLU activation functions, but do not perform batch normalization.

In order to prevent overfitting on the training simulations, we include in the CNN's final architecture significant levels of dropout. This technique randomly sets to zero a percentage of neuron activations during training. Specifically, we use dropout percentages of $p = 0.15$ for each convolutional block and $p = 0.4$ for each fully connected block. Additionally, we introduce a large ℓ_2 penalty term with normalization value $\lambda = 0.0275$ on the network weights. In applying dropout in both the convolutional and fully connected layers, the network is forced to train on a smaller subset of active neurons, leading to underutilization of the network's capacity. Moreover, with the ℓ_2 penalty term in the loss function, the network's flexibility, and subsequently its ability to learn specific features, is limited. While these measures ultimately limit the constraining power of the CNN, they ensure robustness and generalizability, and thus protect against the fact that the SimBIG forward models, and in general any forward model, are approximate.

CNN training is performed using a supervised learning approach. We optimize the weights of the network to minimize the mean-squared-error (MSE) loss between $\hat{\boldsymbol{\theta}}_{\rm normed}$ and $\boldsymbol{\theta}^{\rm true}_{\rm normed}$, where we normalize both $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{\rm true}$ to (0, 1), to prevent their varying ranges from affecting the loss differently. The optimization is performed using stochastic gradient descent with momentum $\beta = 0.9$. The neural network is trained in minibatches of 32 galaxy fields. We use the OneCycleLR learning rate scheduler, which involves gradually increasing and then decreasing the learning rate during a single training cycle and has been shown to lead to faster convergence and improved generalization [66]. We use a maximal learning rate of $r = 0.01$. During training, the input fields are also randomly flipped horizontally and vertically with $p = 0.5$ to further improve network generalization. We train the CNN on a single A100 GPU core until the MSE computed on the validation set has not improved for 20 consecutive epochs. Training the CNN in this context takes roughly 8 h.

The CNN's architecture and hyperparameters are determined through experimentation and are roughly modeled off of previous successful image classifiers in [67,68]. To determine the specifics of our network, we train 60 networks using the Optuna hyperparameter framework [69]. Specifically, we vary the number of convolutional blocks between 3 and 6, the number of fully connected layers between 1 and 6, the base number of channels of the convolutional blocks between 2 and 14, the width of the fully connected layers between 128 and 1024, the dropout in both convolutional and fully connected layers between $p = 0$ and $p = 0.5$, the ℓ_2 penalty between $\lambda = 10^{-4}$ and $\lambda = 10^{-1}$, and the max learning rate between $r = 10^{-5}$ and $r = 10^{-2}$. Ultimately, we aim to maximize the network's ability to extract relevant features from the galaxy density field while maintaining its ability to generalize beyond the SimBIG training simulations. To that end, we select the network configuration that maximizes the network's MSE on the held-out validation models while minimizing the ratio between training MSE and validation MSE. However, in order to pass the validation tests on the out-of-distribution TEST1 and TEST2, we found that it was necessary to impose slightly stricter regularization on the network. Thus, the dropout and ℓ_2 terms were increased through trial and error from the Optuna output to their reported values. Ultimately, the significant amounts of regularization are included due to the model's tendency to overfit on the relatively small dataset size.

B. Weight marginalization

In order to further prevent the CNN from overfitting on the training set, we perform a weight marginalization step, converting our CNN into a Bayesian neural network (BNN). In contrast to other neural networks, BNNs train the model weights as a distribution rather than searching for an optimal value. This allows them to capture the uncertainty in the weights and outputs of the model. The ultimate goal of BNNs is to quantify the uncertainty introduced by the models in terms of outputs and weights so as to explain the trustworthiness of the prediction.

In this work, we use stochastic weight averaging (SWA) [70,71]. SWA is predicated on the observation that the parameters of deep neural networks often converge to the edges of low-loss regions. This edge-type convergence is suboptimal, as these solutions are more susceptible to the shift between train and test error surfaces. SWA approximates the posterior distribution of the weights of the CNN as a normal distribution, whose mean and covariance are given by

$$\bar{w} = \frac{1}{N_{\rm swa}} \sum_{n=1}^{N_{\rm swa}} w_n, \quad \Sigma = \frac{1}{N_{\rm swa}} \sum_{n=1}^{N_{\rm swa}} (w_n - \bar{w})(w_n - \bar{w})^T, \tag{3.1}$$

respectively, where w are the weights of the network, n is the time step during network optimization/training, and $N_{\rm swa}$ are the total steps over which SWA is performed.

By adopting this scheme, SWA solutions tend to converge to the center of flat loss regions, thereby leading to more stable and generalizable solutions. Indeed, SWA has already been shown to lead to better generalization to out-of-distribution data [71], which is expected to improve the robustness of our analysis. Moreover, SWA has been shown to outperform competing methods in multiple tasks [70] and has been previously applied to astrophysics [72] and cosmology [73]. We use the publicly available cosmoSWAG implementation [74]. The compressed galaxy field that we feed as input to SBI is the output of the SWA network: a set of ten samples of the posterior distribution weights of the CNN—a 50-dimensional data vector.

C. Simulation-based inference

After training the CNN, we use the SimBIG SBI framework to estimate posterior distributions of the cosmological parameters $\boldsymbol{\theta}$ from the compressed representation of the observables obtained from the CNN $\hat{\boldsymbol{\theta}}$. We represent this posterior as $p(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$.

There are multiple existing frameworks for SBI, such as approximate Bayesian computation (e.g., [75–79]), neural ratio estimation (e.g., [80–84]), neural likelihood estimation (e.g., [85–87]), and neural posterior score estimation (e.g., [88,89]). We use neural posterior estimation (e.g., [90–94]), which uses a neural density estimator (NDE) to estimate the posterior distribution from a training set. In this case, the training set consists of the ground-truth/CNN-compressed $\{\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\}$ parameter pairs of the SimBIG forward models. We use the publicly available SBI implementation from Tejero-Cantero *et al.* [95].

Previous SimBIG analyses employed a masked autoregressive flow [96] as the density estimator. For our density estimator, we instead use neural spline flows (NSFs) [97], a more expressive alternative. Denoting our NSF as $q_\phi(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$, where ϕ represents its hyperparameters, we train q_ϕ by minimizing the kullback-leibler divergence between $p(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ and $q_\phi(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}})$. This is equivalent to maximizing the log-likelihood over the training set of SimBIG forward models. In practice, we split the catalogs into a training and validation set with 90/10 split and use an early stopping procedure to prevent overfitting by stopping training when the validation log-likelihood has failed to increase after 20 epochs. Additionally, to improve the robustness of our NDE, we use an ensemble of five NSFs, which has been shown to produce more reliable approximations [98,99].

D. Validation

Before analyzing observations, we first validate our posterior estimation in two stages. First, we validate on the 5180 simulations that were excluded from the training of our pipeline. We refer to this as the "NDE accuracy test." Second, we conduct the SimBIG "mock challenge," where we validate our analysis on the suite of test simulations described in Sec. II.

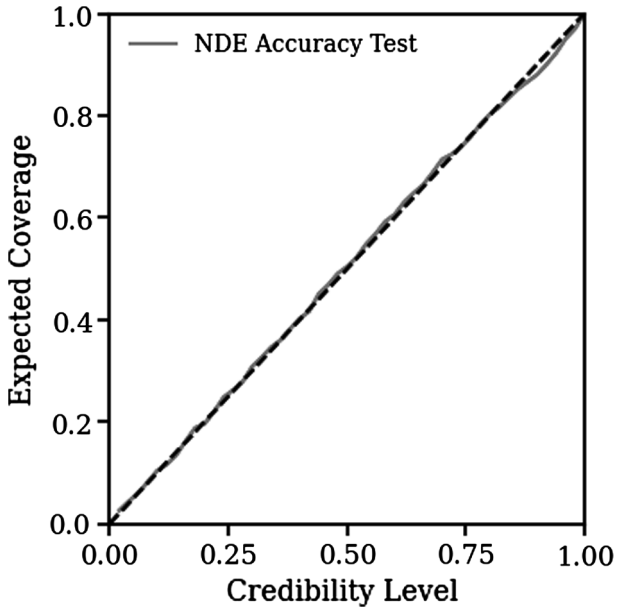


FIG. 2. TARP expected coverage probability vs probability level. For an accurate posterior estimator, the line will follow the diagonal, while deviations from the diagonal are indicative of over- or underconfidence. We show the NDE accuracy test, using 5180 of our base simulations that were not used in the training of our CNN.

For the NDE accuracy test, we use the tests of accuracy with random point (TARP) expected coverage probability (ECP) test as our metric. ECP is a necessary and sufficient test for the optimality of the estimated posterior, q_ϕ [100,101]. $p(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) \equiv q_\phi(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ is only true in the limit of infinite data, and therefore we can only test for approximate equality, which is satisfied if and only if

$$\mathrm{ECP}(\hat{p}, \alpha) = 1 - \alpha \quad \forall \ \alpha \in [0, 1], \tag{3.2}$$

where $\mathrm{ECP}(\hat{p}, \alpha)$ is the expected coverage probability of the posterior estimate $\hat{p}$. TARP coverage probabilities are a robust method for estimating ECP that do not rely on evaluations of the posterior estimate. We can use it to calculate ECP for both the full-dimensional parameter space or for each parameter separately. The latter is equivalent to the simulation-based calibration [102] used in the other SimBIG analyses.

We present the results of our NDE accuracy test using TARP in Fig. 2, where we plot the ECP versus the confidence level $1 - \alpha$.[5] We evaluate the TARP ECP over the full dimensionality of our parameter space. If the ECP and confidence level are equal for every $\alpha \in [0, 1]$,

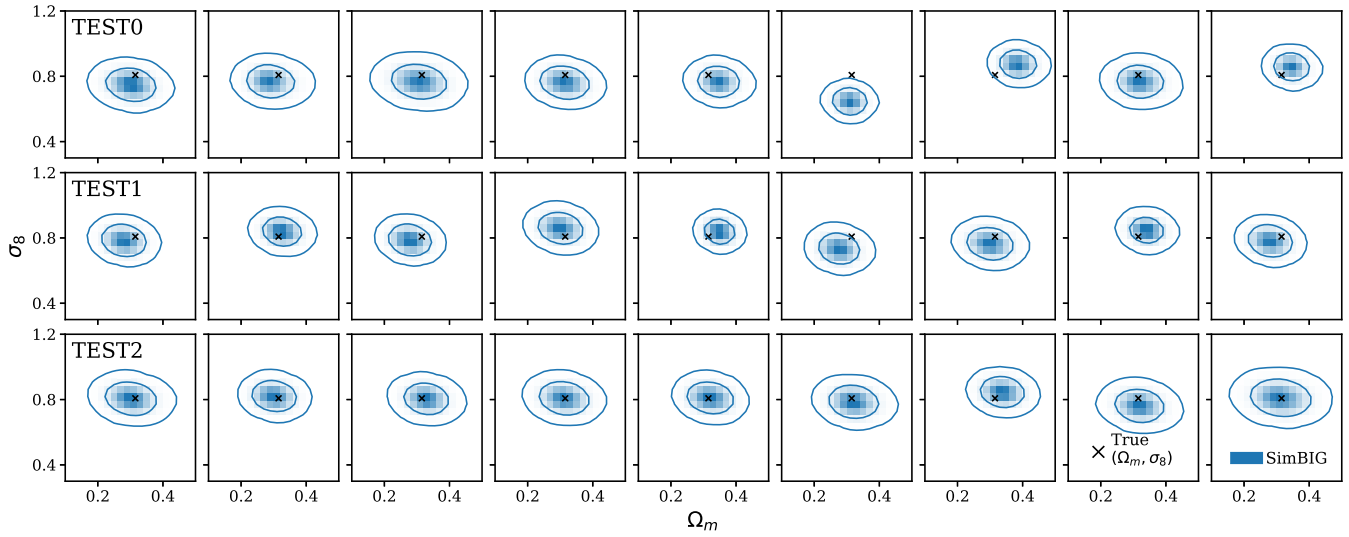
[5]This figure is often referred to as a probability-probability plot.

i.e., it follows a diagonal line, then the estimator is well calibrated since the probability of our posterior estimate containing the true parameter values matches the actual confidence level. We find that the NDE accuracy test is perfectly calibrated, as the ECP line is perfectly in the diagonal.

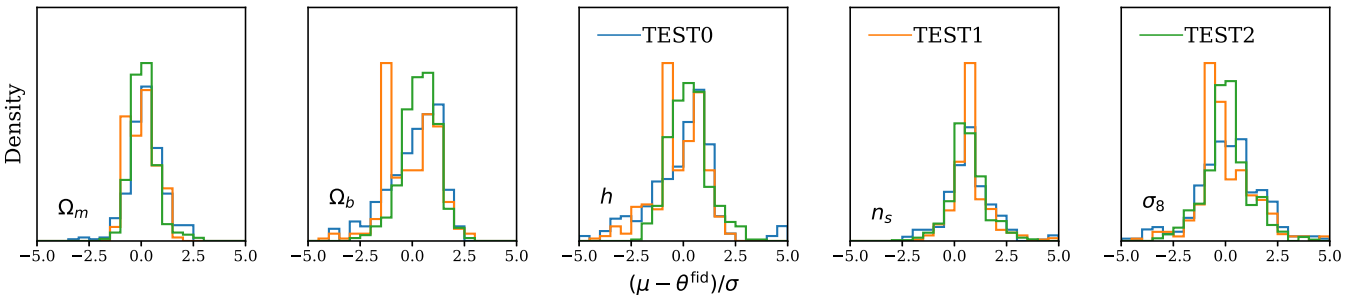
We then move on to the SimBIG mock challenge. Figure 3(a) shows the marginalized two-dimensional posterior distribution of $\{\Omega_m, \sigma_8\}$ for nine randomly selected simulations from each of the test sets: TEST0 (top), TEST1 (center), and TEST2 (bottom). We mark the true Ω_m and σ_8 values in each panel (black x). For all three test sets, the posteriors appear to be well calibrated and unbiased, which qualitatively demonstrate the robustness of our analysis.

Next, we assess robustness more quantitatively using TEST0, TEST1, and TEST2. For the test simulations, we cannot use the same method as the NDE accuracy test due to the fact that the ECP relies on averaging over the prior distribution, but all of these simulations are run at fixed fiducial cosmologies. Therefore, we follow [49] and we assess robustness by comparing the likelihoods over the three test sets. Specifically, we compute the posterior mean μ and standard deviation σ for each ΛCDM parameter for each suite of test simulations. Then, we analyze the difference between μ and the true parameter value $\theta^{\rm fid}$ in units of σ. For a robust pipeline, we expect to find consistency of these estimates across all three datasets. On the other hand, variations between the distributions would be indicative of likelihood variations that come from changing the forward model and imply that our analysis is sensitive to model variations.

In Fig. 3(b), we present the likelihoods of TEST0 (blue), TEST1 (orange), and TEST2 (green) for each of the ΛCDM parameters. We find consistent distributions for all parameters across test sets. This indicates that our posterior inference is robust to variations in the forward model.



(a) Posterior distributions for (Ω_m, σ_8). Each row shows nine randomly selected examples taken from each test set, as labeled. True parameters are marked in black and the contours represent the 68 and 95 percentiles.



(b) Distributions of the differences between the posterior mean μ and the true parameter $\boldsymbol{\theta}^{\rm fid}$ normalized by the posterior standard deviation σ, for each cosmological parameter and test set. Differences between the distributions among the three datasets would be indicative of likelihood variations when we change our forward model. We find good agreement across all three forward models and all five parameters.

FIG. 3. Validation of our model on the SimBIG mock challenge data: (a) Marginalized two-dimensional posterior distribution of $\{\Omega_m, \sigma_8\}$. (b) Likelihoods of ΛCDM parameters.

It also suggests that our use of weight marginalization leads to better generalization properties.

These validation tests form a crucial part of our analysis. We note that it is possible to obtain significantly tighter constraints that pass only the NDE accuracy test. However, in doing so, we would need to assume that our forward model accurately models every aspect of the observations. Given the complexities of galaxy formation, *any* forward model of galaxy clustering is an approximate model. Hence, validating that we can successfully infer unbiased cosmological constraints from simulated test galaxy catalogs generated with different forward models (TEST1 and TEST2) serves as a powerful test against model misspecification, even if it comes at the expense of significant constraining power. In future work, we will explore additional tests of model misspecification and "blind challenges," where we test our analysis on simulations without knowing the true cosmological parameters or the forward model used to generate them.

IV. RESULTS

In Fig. 4, we present the posterior distribution of all ΛCDM cosmological parameters inferred from our field-level analysis of the BOSS CMASS SGC using SimBIG (orange). In the right panels, we focus on the growth of structure parameters Ω_m and σ_8. The diagonal panels present the 1D marginalized posteriors; the rest of the panels present marginalized 2D posteriors of different parameter pairs. The contours represent the 68 and 95 percentiles and the ranges of the panels match the prior. For comparison, we include posteriors from the SimBIG $P_\ell(k_{\rm max} < 0.5\ h/{\rm Mpc})$ analysis (gray) [63], as well as the constraints from the PT based $P_\ell(k_{\rm max} < 0.25\ h/{\rm Mpc})$ analysis of the CMASS SGC sample (dashed) [8].

Overall, our field-level analysis using the CNN provides tighter, yet consistent, cosmological constraints to the previous BOSS analyses. Specifically, our constraints on Ω_m and σ_8 are 1.76× and 1.92× tighter than the SimBIG P_ℓ analysis. Moreover, our constraints on Ω_m are in line with the PT-based P_ℓ analysis, and ours on σ_8 are 2.65× tighter. This higher constraining power is expected. Indeed, by using the full galaxy field, we are able to exploit non-Gaussian cosmological information on nonlinear scales that is inaccessible to P_ℓ analyses. Moreover, in using the SimBIG SBI approach, we are able to more robustly account for observational systematics compared to the standard clustering analyses.

In fact, with the added constraining power of our field-level analysis, we also can place significant constraints on

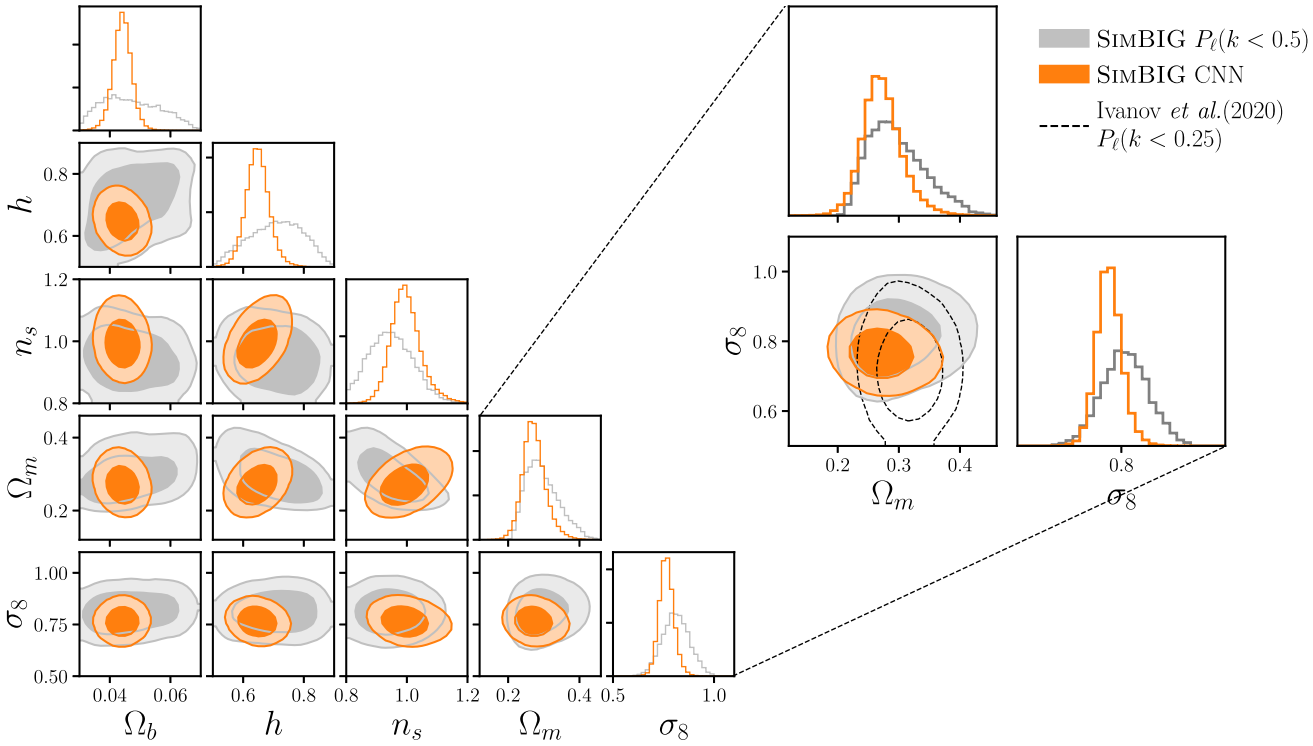


FIG. 4. Left: posterior distributions for all ΛCDM cosmological parameters from our CNN-based field-level inference of BOSS observations (orange). For comparison, we include the SimBIG P_ℓ analysis (gray). The contours represent the 68% and 95% confidence intervals. Our CNN-based field-level inference produces tighter, yet consistent, constraints to the SimBIG P_ℓ analysis. Right: posterior distributions for Ω_m and σ_8. For comparison, we include posteriors from the SimBIG P_ℓ analysis (gray) and the standard PT-based P_ℓ analysis (black dashed) [8]. Our analysis constrains Ω_m and σ_8 1.76 and 1.92× tighter than the SimBIG P_ℓ analysis. Moreover, our constraints on Ω_m are in line with the PT-based P_ℓ analysis, and ours on σ_8 are 2.65× tighter.

$H_0 = 63.1 \pm 4.1$ km/s/Mpc, albeit weaker than those on Ω_m and σ_8. This is in contrast to standard P_ℓ analyses, which cannot independently constrain H_0 and typically rely on priors from big bang nucleosynthesis or cosmic microwave background experiments. Our constraints support a low value of H_0 in good agreement with *Planck* constraints [103]. However, we do not have enough constraining power to make strong statements. We will further investigate the cosmological implications of this result and how they compare with other surveys and cosmological probes in an accompanying paper [104].

V. CONCLUSIONS

In this paper, we present cosmological constraints from a field-level analysis of the CMASS galaxy catalogs using simulation-based inference. We demonstrate that our analysis passes a number of stringent validation tests, including a robustness test based on simulations constructed using different forward models. These test sets provide key validation against model misspecification and demonstrate some robustness against discrepancies between observations and our forward model. It is important to point out, however, that in an ideal situation we would have had a separate set of test simulations that we did not use for calibration.

Furthermore, we show that our cosmological parameter constraints are consistent but significantly tighter than those from P_ℓ analyses. In particular, our constraints on Ω_m and σ_8 are in line and 2.65× tighter than the standard PT-based P_ℓ analyses. We are even able to produce significant constraints on H_0, without any priors from external experiments. These improvements demonstrate that our method successfully extracts additional non-Gaussian and nonlinear cosmological information from the galaxy distribution.

As simulations become more realistic and efficient in the future, we will be able to extend our analyses to smaller scales at the larger volumes covered by upcoming surveys such as the Dark Energy Spectroscopic Instrument [105–107], Subaru Prime Focus Spectrograph [108,109], the ESA *Euclid* satellite mission [110], and the Nancy Grace Roman space telescope [111,112]. Our results demonstrate that these analyses will be able to produce leading cosmological constraints from galaxy clustering. The methodology and tests presented in this paper lay the groundwork for such analyses.

In accompanying papers [104,113], we present the SimBIG analysis of galaxy clustering using two summary statistics: the galaxy bispectrum and the wavelet scattering transform statistics. Furthermore, in [104], we present a comparison of the different SimBIG analyses, including the field-level constraints presented in this work. We also discuss their cosmological implications and present forecasts for extending SimBIG to upcoming galaxy surveys.

Observational data used in this paper can be found at [48]. The cosmoSWAG implementation and TARP are publicly available at GitHub [74,101], respectively.

ACKNOWLEDGMENTS

It is a pleasure to thank Mikhail M. Ivanov for providing us with the posteriors used for comparison and Ben Wandelt for discussions that greatly helped the papers. We thank the Learning the Universe Collaboration for helpful feedback and stimulating discussions. P. L. acknowledges support from the Simons Foundation. J. H. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101025187. A. M. D. acknowledges funding from Tomalla Foundation for Research in Gravity.

[1] N. Kaiser, Mon. Not. R. Astron. Soc. **227**, 1 (1987).
[2] A. Hamilton, Mon. Not. R. Astron. Soc. **289**, 285 (1997).
[3] J. A. Peacock, S. Cole, P. Norberg, C. M. Baugh, J. Bland-Hawthorn, T. Bridges, R. D. Cannon, M. Colless, C. Collins, W. Couch *et al.*, Nature (London) **410**, 169 (2001).
[4] E. Hawkins, S. Maddox, S. Cole, O. Lahav, D. S. Madgwick, P. Norberg, J. A. Peacock, I. K. Baldry, C. M. Baugh, J. Bland-Hawthorn *et al.*, Mon. Not. R. Astron. Soc. **346**, 78 (2003).
[5] M. Tegmark, D. J. Eisenstein, M. A. Strauss, D. H. Weinberg, M. R. Blanton, J. A. Frieman, M. Fukugita, J. E. Gunn, A. J. Hamilton, G. R. Knapp *et al.*, Phys. Rev. D **74**, 123507 (2006).
[6] L. Guzzo, M. Pierleoni, B. Meneux, E. Branchini, O. Le Fèvre, C. Marinoni, B. Garilli, J. Blaizot, G. De Lucia, A. Pollo *et al.*, Nature (London) **451**, 541 (2008).
[7] F. Beutler *et al.*, Mon. Not. R. Astron. Soc. **466**, 2242 (2017).
[8] M. M. Ivanov, M. Simonović, and M. Zaldarriaga, Phys. Rev. D **101**, 083504 (2020).
[9] Y. Kobayashi, T. Nishimichi, M. Takada, and H. Miyatake, Phys. Rev. D **105**, 083517 (2022).
[10] F. Bernardeau, S. Colombi, E. Gaztanaga, and R. Scoccimarro, Phys. Rep. **367**, 1 (2002).
[11] V. Desjacques, D. Jeong, and F. Schmidt, Phys. Rep. **733**, 1 (2018).

[12] H. Gil-Marín, W. J. Percival, L. Verde, J. R. Brownstein, C.-H. Chuang, F.-S. Kitaura, S. A. Rodríguez-Torres, and M. D. Olmstead, Mon. Not. R. Astron. Soc. **465**, 1757 (2017).

[13] G. D'Amico, Y. Donath, M. Lewandowski, L. Senatore, and P. Zhang, arXiv:2206.08327.

[14] O. H. Philcox and M. M. Ivanov, Phys. Rev. D **105**, 043517 (2022).

[15] M. M. Ivanov, O. H. E. Philcox, G. Cabass, T. Nishimichi, M. Simonović, and M. Zaldarriaga, Phys. Rev. D **107**, 083515 (2023).

[16] C. Hahn, F. Villaescusa-Navarro, E. Castorina, and R. Scoccimarro, J. Cosmol. Astropart. Phys. 03 (2020) 040.

[17] C. Hahn and F. Villaescusa-Navarro, J. Cosmol. Astropart. Phys. (2021) 029.

[18] E. Massara, F. Villaescusa-Navarro, C. Hahn, M. M. Abidi, M. Eickenberg, S. Ho, P. Lemos, A. Moradinezhad Dizgah, and B. Régaldo-Saint Blancard, Astrophys. J. **951**, 70 (2023).

[19] Y. Wang, G.-B. Zhao, K. Koyama, W. J. Percival, R. Takahashi, C. Hikage, H. Gil-Marín, C. Hahn, R. Zhao, W. Zhang, X. Mu, Y. Yu, H.-M. Zhu, and F. Ge, arXiv:2202.05248.

[20] J. Hou, A. Moradinezhad Dizgah, C. Hahn, and E. Massara, J. Cosmol. Astropart. Phys. 03 (2023) 045.

[21] M. Eickenberg, E. Allys, A. Moradinezhad Dizgah, P. Lemos, E. Massara, M. Abidi, C. Hahn, S. Hassan, B. Regaldo-Saint Blancard, S. Ho, S. Mallat, J. Andén, and F. Villaescusa-Navarro, arXiv:2204.07646.

[22] C. Hahn, R. Scoccimarro, M. R. Blanton, J. L. Tinker, and S. A. Rodríguez-Torres, Mon. Not. R. Astron. Soc. **467**, 1940 (2017).

[23] D. Bianchi, A. Burden, W. J. Percival, D. Brooks, R. N. Cahn, J. E. Forero-Romero, M. Levi, A. J. Ross, and G. Tarle, Mon. Not. R. Astron. Soc. **481**, 2338 (2018).

[24] A. J. Ross, W. J. Percival, A. G. Sánchez, L. Samushia, S. Ho, E. Kazin, M. Manera, B. Reid, M. White, R. Tojeiro *et al.*, Mon. Not. R. Astron. Soc. **424**, 564 (2012).

[25] A. J. Ross *et al.*, Mon. Not. R. Astron. Soc. **464**, 1168 (2017).

[26] R. Scoccimarro, Astrophys. J. **542**, 1 (2000).

[27] E. Sellentin and A. F. Heavens, Mon. Not. R. Astron. Soc. **473**, 2355 (2018).

[28] C. Hahn, F. Beutler, M. Sinha, A. Berlind, S. Ho, and D. W. Hogg, Mon. Not. R. Astron. Soc. **485**, 2956 (2019).

[29] C. Hahn, M. Vakili, K. Walsh, A. P. Hearin, D. W. Hogg, and D. Campbell, Mon. Not. R. Astron. Soc. **469**, 2791 (2017).

[30] L. Perreault Levasseur, Y. D. Hezaveh, and R. H. Wechsler, Astrophys. J. Lett. **850**, L7 (2017).

[31] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **127**, 241103 (2021).

[32] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, Mon. Not. R. Astron. Soc. **488**, 4440 (2019).

[33] S. Wagner-Carena, J. W. Park, S. Birrer, P. J. Marshall, A. Roodman, and R. H. Wechsler (LSST Dark Energy Science Collaboration), Astrophys. J. **909**, 187 (2021).

[34] R. Legin, Y. Hezaveh, L. P. Levasseur, and B. Wandelt, arXiv:2112.05278.

[35] A. Coogan, K. Karchev, and C. Weniger, in *34th Conference on Neural Information Processing Systems* (2020), arXiv:2010.07032.

[36] N. A. Montel, A. Coogan, C. Correa, K. Karchev, and C. Weniger, Mon. Not. R. Astron. Soc. **518**, 2746 (2022).

[37] A. Coogan, N. Anau Montel, K. Karchev, M. W. Grootes, F. Nattino, and C. Weniger, Mon. Not. R. Astron. Soc. **527**, 66 (2024).

[38] J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe, and K. Cranmer, Astrophys. J. **886**, 49 (2019).

[39] S. Mishra-Sharma and K. Cranmer, Phys. Rev. D **105**, 063017 (2022).

[40] K. Karchev, R. Trotta, and C. Weniger, Mon. Not. R. Astron. Soc. **520**, 1056 (2023).

[41] J. Hermans, N. Banik, C. Weniger, G. Bertone, and G. Louppe, Mon. Not. R. Astron. Soc. **507**, 1999 (2021).

[42] K. Karchev, N. Anau Montel, A. Coogan, and C. Weniger, in *36th Conference on Neural Information Processing Systems* (2022), arXiv:2211.04365.

[43] P. Lemos, N. Jeffrey, L. Whiteway, O. Lahav, I. Noam Libeskind, and Y. Hoffman, Phys. Rev. D **103**, 023009 (2021).

[44] C. Hahn and P. Melchior, arXiv:2203.07391.

[45] C. Hahn, M. Eickenberg, S. Ho, J. Hou, P. Lemos, E. Massara, C. Modi, A. M. Dizgah, B. R.-S. Blancard, and M. M. Abidi, arXiv:2211.00660.

[46] D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. A. Prieto, S. F. Anderson, J. A. Arns, É. Aubourg, S. Bailey, E. Balbinot *et al.*, Astron. J. **142**, 72 (2011).

[47] K. S. Dawson, D. J. Schlegel, C. P. Ahn, S. F. Anderson, É. Aubourg, S. Bailey, R. H. Barkhouser, J. E. Bautista, A. Beifiori, A. A. Berlind *et al.*, Astron. J. **145**, 10 (2012).

[48] https://data.sdss.org/sas/dr12/boss/lss/.

[49] C. Hahn, M. Eickenberg, S. Ho, J. Hou, P. Lemos, E. Massara, C. Modi, A. Moradinezhad Dizgah, B. Régaldo-Saint Blancard, and M. M. Abidi, J. Cosmol. Astropart. Phys. 04 (2023) 010.

[50] F. Villaescusa-Navarro, C. Hahn, E. Massara, A. Banerjee, A. M. Delgado, D. K. Ramanah, T. Charnock, E. Giusarma, Y. Li, E. Allys *et al.*, Astrophys. J. Suppl. Ser. **250**, 2 (2020).

[51] P. S. Behroozi, R. H. Wechsler, and H.-Y. Wu, Astrophys. J. **762**, 109 (2012).

[52] Z. Zheng, A. L. Coil, and I. Zehavi, Astrophys. J. **667**, 760 (2007).

[53] S. More, H. Miyatake, M. Takada, B. Diemer, A. V. Kravtsov, N. K. Dalal, A. More, R. Murata, R. Mandelbaum, E. Rozo *et al.*, Astrophys. J. **825**, 39 (2016).

[54] M. Vakili and C. Hahn, Astrophys. J. **872**, 115 (2019).

[55] A. R. Zentner, A. Hearin, F. C. van den Bosch, J. U. Lange, and A. S. Villarreal, Mon. Not. R. Astron. Soc. **485**, 1196 (2019).

[56] B. Hadzhiyska, S. Liu, R. S. Somerville, A. Gabrielpillai, S. Bose, D. Eisenstein, and L. Hernquist, Mon. Not. R. Astron. Soc. **508**, 698 (2021).

[57] J. Carlson and M. White, Astrophys. J. Suppl. Ser. **190**, 311 (2010).
[58] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. White, Astrophys. J. **292**, 371 (1985).
[59] N. A. Maksimova, L. H. Garrison, D. J. Eisenstein, B. Hadzhiyska, S. Bose, and T. P. Satterthwaite, Mon. Not. R. Astron. Soc. **508**, 4017 (2021).
[60] B. Hadzhiyska, D. Eisenstein, S. Bose, L. H. Garrison, and N. Maksimova, Mon. Not. R. Astron. Soc. **509**, 501 (2022).
[61] C. K. Birdsall and D. Fuss, J. Comput. Phys. **3**, 494 (1969).
[62] L. Anderson, E. Aubourg, S. Bailey, D. Bizyaev, M. Blanton, A. S. Bolton, J. Brinkmann, J. R. Brownstein, A. Burden, A. J. Cuesta *et al.*, Mon. Not. R. Astron. Soc. **427**, 3435 (2012).
[63] C. Hahn, M. Abidi, M. Eickenberg, S. Ho, P. Lemos, E. Massara, A. Moradinezhad Dizgah, and B. Régaldo-Saint Blancard, Mach. Learn. Astrophys. 24 (2022).
[64] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, J. Big Data **8**, 1 (2021).
[65] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, Adv. Neural Inf. Process. Syst. **31** (2018).
[66] L. N. Smith and N. Topin, in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (SPIE, 2019), Vol. 11006, pp. 369–386.
[67] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2818–2826.
[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Commun. ACM **60**, 84 (2017).
[69] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019), pp. 2623–2631.
[70] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, Adv. Neural Inf. Process. Syst. **32** (2019).
[71] A. G. Wilson and P. Izmailov, Adv. Neural Inf. Process. Syst. **33**, 4697 (2020).
[72] M. Cranmer, D. Tamayo, H. Rein, P. Battaglia, S. Hadden, P. J. Armitage, S. Ho, and D. N. Spergel, Proc. Natl. Acad. Sci. U.S.A. **118**, e2026053118 (2021).
[73] P. Lemos, M. Cranmer, M. Abidi, C. Hahn, M. Eickenberg, E. Massara, D. Yallup, and S. Ho, Mach. Learn. **4**, 01LT01 (2023).
[74] https://github.com/Pablo-Lemos/cosmoSWAG.
[75] D. B. Rubin, Ann. Stat. **12**, 1151 (1984).
[76] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, Mol. Biol. Evol. **16**, 1791 (1999).
[77] M. A. Beaumont, W. Zhang, and D. J. Balding, Genetics **162**, 2025 (2002).
[78] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré, Proc. Natl. Acad. Sci. U.S.A. **100**, 15324 (2003).
[79] P. Fearnhead and D. Prangle, J. R. Stat. Soc. **74**, 419 (2012).
[80] K. Cranmer, J. Pavez, and G. Louppe, arXiv:1506.02169.
[81] O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann, Bayesian Anal. **17**, 1 (2022).
[82] J. Hermans, V. Begy, and G. Louppe, in *International Conference on Machine Learning* (PMLR, 2020), pp. 4239–4248.
[83] C. Durkan, I. Murray, and G. Papamakarios, in *International Conference on Machine Learning* (PMLR, 2020), pp. 2771–2781.
[84] B. K. Miller, C. Weniger, and P. Forré, arXiv:2210.06170.
[85] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott, J. Comput. Graph. Stat. **27**, 1 (2018).
[86] G. Papamakarios, D. Sterratt, and I. Murray, in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), pp. 837–848.
[87] D. T. Frazier, D. J. Nott, C. Drovandi, and R. Kohn, J. Am. Stat. Assoc. **118**, 1 (2022).
[88] L. Sharrock, J. Simons, S. Liu, and M. Beaumont, arXiv:2210.04872.
[89] T. Geffner, G. Papamakarios, and A. Mnih, arXiv:2209.14249.
[90] D. Rezende and S. Mohamed, in *International Conference on Machine Learning* (PMLR, 2015), pp. 1530–1538.
[91] G. Papamakarios and I. Murray, Adv. Neural Inf. Process. Syst. **29** (2016).
[92] J.-M. Lueckmann, G. Bassetto, T. Karaletsos, and J. H. Macke, Proc. Mach. Learn. Res. **96**, 32 (2019).
[93] J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, Adv. Neural Inf. Process. Syst. **30** (2017).
[94] D. Greenberg, M. Nonnenmacher, and J. Macke, in *International Conference on Machine Learning* (PMLR, 2019), pp. 2404–2414.
[95] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke, J. Open Source Softwaare **5**, 2505 (2020).
[96] G. Papamakarios, T. Pavlakou, and I. Murray, arXiv:1705.07057.
[97] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Adv. Neural Inf. Process. Syst. **32** (2019).
[98] B. Lakshminarayanan, A. Pritzel, and C. Blundell, Adv. Neural Inf. Process. Syst. **30** (2017).
[99] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, V. Begy, and G. Louppe, stat **1050**, 4 (2022), https://openreview.net/forum?id=LHAbHkt6Aq.
[100] P. Lemos, A. Coogan, Y. Hezaveh, and L. Perreault-Levasseur, arXiv:2302.03026.
[101] https://github.com/Ciela-Institute/tarp.
[102] S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman, arXiv:1804.06788.
[103] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. Banday, R. Barreiro, N. Bartolo, S. Basak *et al.*, Astron. Astrophys. **641**, A6 (2020).
[104] ChangHoon Hahn, Michael Eickenberg, Shirley Ho, Jiamin Hou, Pablo Lemos *et al.* (SimBIG Collaboration), this issue, Phys. Rev. D **109**, 083534 (2024).
[105] A. Aghamousa *et al.* (DESI Collaboration), arXiv:1611.00036.
[106] A. Aghamousa *et al.* (DESI Collaboration), arXiv:1611.00037.

[107] B. Abareshi, Astron. J. **164**, 207 (2022).
[108] M. Takada *et al.*, Publ. Astron. Soc. Jpn. **66**, R1 (2014).
[109] N. Tamura *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **9908**, 99081M (2016).
[110] R. Laureijs *et al.*, arXiv:1110.3193.
[111] D. Spergel *et al.*, arXiv:1503.03757.
[112] Y. Wang, Z. Zhai, A. Alavi, E. Massara, A. Pisani, A. Benson, C. M. Hirata, L. Samushia, D. H. Weinberg, J. Colbert, O. Doré, T. Eifler, C. Heinrich, S. Ho, E. Krause, N. Padmanabhan, D. Spergel, and H. I. Teplitz, Astrophys. J. **928**, 1 (2022).
[113] Bruno Régaldo-Saint Blancard, ChangHoon Hahn, Shirley Ho, Jiamin Hou, Pablo Lemos *et al.* (SimBIG Collaboration), preceding paper, Phys. Rev. D **109**, 083535 (2024).