Field-level cosmological model selection: field-level simulation-based inference for Stage IV cosmic shear can distinguish dynamical dark energy

Alessio Spurio Mancini,[1,*] Kiyam Lin,[2] and Jason D. McEwen[2]

[1]*Department of Physics, Royal Holloway, University of London, Egham Hill, Egham, UK*
[2]*Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, Surrey, RH5 6NT, UK*

(Dated: October 14, 2024)

We present a framework that for the first time allows Bayesian model comparison to be performed for field-level inference of cosmological models. We achieve this by taking a simulation-based inference (SBI) approach using neural likelihood estimation, which we couple with the learned harmonic mean estimator in order to compute the Bayesian evidence for model comparison. We apply our framework to mock Stage IV cosmic shear observations to assess its effectiveness at distinguishing between various models of dark energy. If the recent DESI results that provided exciting hints of dynamical dark energy were indeed the true underlying model, our analysis shows Stage IV cosmic shear surveys could definitively detect dynamical dark energy. We also perform traditional power spectrum likelihood-based inference for comparison, which we find is not able to distinguish between dark energy models, highlighting the enhanced constraining power for model comparison of our field-level SBI approach.

I. INTRODUCTION

Central questions in cosmology are often those of model comparison. For example, what model best describes the underlying nature of dark energy? The concordance ΛCDM model attributes dark energy to Einstein's cosmological constant Λ. In the wCDM model, the dark energy equation-of-state parameter w is constant in time but allowed to vary from -1 to be constrained by observations. In dynamical dark energy models, such as the $w_0 w_a$CDM model, w is free to evolve over time. Recent results from the DESI collaboration [1] hint at the exciting possibility of dynamical dark energy, although the consensus at present is that there is no evidence to prefer a model more complicated than ΛCDM. In this work we present a cosmic shear analysis pipeline for cosmological model comparison and study its effectiveness at comparing the ΛCDM, wCDM and $w_0 w_a$CDM cosmological models.

While traditional cosmological inference pipelines are typically based on likelihood-based analysis of two-point statistics, it is widely known that probes of the large-scale structure contain a great deal of cosmological information beyond two-point statistics due to the non-linear nature of gravity. Field-level inference is capable of capturing this higher-order statistical information [*e.g.* 2–8]). Upcoming Stage IV surveys of the large-scale structure, such as *Euclid* [9], Rubin Observatory Legacy Survey of Space and Time (Rubin-LSST) [10] or *Roman* [11] will acquire data that contain significant high-order cosmological information in the observed fields.

However, field-level inference with such data is challenging due to the high-dimensional nature of the parameter space to be inferred and the complexity of the forward model. Typically only parameter estimation is considered, as model comparison is too computationally costly. Yet, the more information we acquire, the better we can distinguish between underlying models. Consequently, performing model comparison on field-level data may enable us to definitively determine which dark energy model best describes our Universe.

An alternative to the aforementioned likelihood-based field-level inference approaches are those that employ simulation-based inference (SBI) [*e.g.* 12]. In this paradigm it is possible to run forward simulations that are able to fully propagate all known uncertainties from parameters to data without needing to explicitly define their corresponding probability distributions. Thus, this approach captures all uncertainties in the data without any statistical simplifications. Modern SBI approaches based on neural density estimation have been applied successfully to two-point statistical analyses in cosmology [13–16] and to field-level analyses [17–21]. Since SBI methods still require large numbers of simulations for training, accelerating simulations is important and becomes even more pertinent for field-level analyses. In particular, neural emulators, such as `CosmoPower` [22], can offer considerable computational savings that in many cases are essential. Furthermore, they also typically support automatic differentiation, which can be leveraged for further acceleration or to reduce the volume of training data needed [23, 24].

Bayesian model comparison provides a principled framework to distinguish between models—naturally incorporating Occam's razor to trade off model complexity and goodness of fit—that has already found widespread use in cosmology [25]. Model comparison requires computation of the Bayesian evidence, which is computationally challenging even in moderate numbers of dimensions. Nested sampling [26] is often used to compute the evidence, as implemented in numerous algorithms [*e.g.* 27–34], although this requires coupling sampling and evidence calculation. Recently, the learned harmonic mean estimator [35] has been presented as an alternative to nested sampling that is flexible, robust, and scalable [36–38]. Moreover, as the learned harmonic mean requires posterior samples only, it is agnostic to the sampling strategy adopted and so can be combined with accelerated sampling techniques, such as the No U-Turn Sampler (NUTS; [39]), `FlowMC` [40], or others, as demonstrated already [38, 41]. However, Bayesian model comparison has not typically been considered for field-

[*] alessio.spuriomancini@rhul.ac.uk

level analyses due to the high-dimensional parameter spaces involved and the difficulties in scaling evidence computation to those dimensions (a notable exception is proximal nested sampling [33, 42], which has been scaled to field-level inference but that is restricted to convex likelihoods and so not applicable to complex cosmological models).

While field-level SBI approaches capture high-order statistical information from the field, the underlying parameter space includes parameters of interest only. In contrast, likelihood-based field-level approaches consider the pixels of the observed or initial field as parameters to be inferred, resulting in very high dimensional parameter spaces. The reduced parameter dimension of field-level SBI opens up the possibility of cosmological model comparison for field-level inference. Model comparison for modern neural SBI approaches was first considered by Ref. [43], where the flexibility of the learned harmonic mean estimator was exploited. Alternative approaches to estimate the evidence that are applicable for SBI have since been introduced where a model is trained specifically to compute the evidence [44, 45].

In this work we present a framework for field-level Bayesian inference, where for the first time we consider not only parameter estimation but also model selection. This is achieved by performing field-level SBI, specifically neural likelihood estimation (NLE; [46]), where cosmological forward models are accelerated by the `CosmoPower` emulator [22, 47], that we couple with the learned harmonic mean for cosmological model comparison [35, 37, 43]. We demonstrate a field-level pipeline on simulated cosmic shear observations, showing that Stage IV surveys can distinguish between different models of dark energy. For comparison, we also consider a likelihood-based analysis based on two-point statistics and demonstrate that it is not able to distinguish between different models, emphasising the effectiveness of field-level inference.

The remainder of this article is structured as follows. Section II details the methodology introduced for field-level SBI that also supports cosmological model comparison. In Section III we apply our framework to mock Stage IV cosmic shear observations to assess its effectiveness at distinguishing between various models of dark energy. Concluding remarks are made in Section IV.

II. METHODOLOGY

We introduce a framework for field-level inference that also supports cosmological model comparison. Specifically, we consider an SBI approach based on NLE (neural likelihood estimation) that we couple with the learned harmonic mean estimator. First, however, we outline the traditional power spectrum likelihood-based inference approach that we consider for comparison.

A. Power spectrum likelihood-based inference

For comparison purposes, we perform a likelihood-based analysis of the weak lensing shear power spectrum. The likelihood is assumed to be Gaussian, following the setup presented in Ref. [7]. The log-likelihood is given by

$$\log p(\boldsymbol{d}|\boldsymbol{\theta}) = -\frac{1}{2}[\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})]^T \boldsymbol{C}^{-1}[\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})], \qquad (1)$$

up to a constant, where $\boldsymbol{\theta}$ represents the underlying cosmological parameters, $\boldsymbol{d}$ is the data vector and $\boldsymbol{C}$ is the covariance matrix for a fixed fiducial cosmology (the same used to generate the mock data vector). The theory shear power spectrum $\boldsymbol{\mu}(\boldsymbol{\theta})$ is calculated from the underlying matter power spectrum using `jax-cosmo`[1] [48]. The non-linear matter power spectrum is provided by `CosmoPower-JAX`[2] [47], a JAX implementation of `CosmoPower` [22], which provides a neural network to emulate the non-linear matter power spectrum. In this work we couple `CosmoPower-JAX` with `jax-cosmo`, using the former to emulate the non-linear prescription given by `HMCode` [49]. `HMCode` provides a parameterised prescription to account for baryonic feedback; in this analysis we fix the baryonic parameters c_{min} and η_0 to their dark matter-only values, 3.13 and 0.603, respectively.

For the simulated data, we make use of a modified version of `sbi_lens`[3] [7] (adding support for the $w_0 w_a$CDM model) to generate correlated convergence maps following a log-normal prescription with Gaussian noise across five tomographic bins. The simulated data is configured to approximately mimic a Stage IV survey. We use `lenstool`[4] [50] to calculate the auto and cross power spectra from the simulated noisy convergence maps.

To generate posterior samples we perform Markov chain Monte Carlo (MCMC) sampling using the NUTS [39] sampler implemented in the `NumPyro`[5] differentiable probabilistic programming library [51, 52].

B. Field-level SBI inference

To perform field-level SBI we do not need an analytical prescription of the likelihood. Instead, we take an NLE (neural likelihood estimation; [46]) approach and learn an implicit likelihood from forward simulated data-parameter pairs. In contrast to Ref. [7] who adopt neural posterior estimation (NPE; [53–55]), we adopt NLE since it provides greater flexibility in the choice of proposal used for generating training data and the neural density estimator can be integrated within an MCMC framework that provides statistical guarantees. Moreover, it simplifies evidence calculation with the learned harmonic mean [43].[6]

[1] https://github.com/DifferentiableUniverseInitiative/jax_cosmo

[2] https://github.com/dpiras/cosmopower-jax

[3] https://github.com/DifferentiableUniverseInitiative/sbi_lens

[4] https://github.com/apetri/LensTools

[5] https://github.com/pyro-ppl/numpyro

[6] The learned harmonic mean requires the evaluation of the likelihood, or its surrogate, at posterior samples, hence model comparison with NPE requires two density estimator to be trained [43].

NLE involves training a conditional density estimator $q_\phi(\boldsymbol{d}|\boldsymbol{\theta})$ to act as a surrogate for the likelihood $p(\boldsymbol{d}|\boldsymbol{\theta})$ (considering it as a probability distribution over the data), where ϕ represent the parameters of the density estimator (i.e. neural network weights). Given paired training data $\{\boldsymbol{\theta}_i, \boldsymbol{d}_i\}$ for parameters drawn from an arbitray proposal $\boldsymbol{\theta}_i \sim \tilde{p}(\boldsymbol{\theta})$, the NLE density estimator can be trained by minimising the negative log-likelihood of the surrogate, which can be shown to be equivalent to maximising the Kullback-Leibler divergence $D_{\mathrm{KL}}(\cdot\|\cdot)$ between the likelihood and its surrogate:

$$\mathbb{E}_{p(\boldsymbol{d}|\boldsymbol{\theta})\tilde{p}(\boldsymbol{\theta})}\left[-\log q_\phi(\boldsymbol{d}|\boldsymbol{\theta})\right] = \mathbb{E}_{\tilde{p}(\theta)}\left[D_{\mathrm{KL}}(p(\boldsymbol{d}|\boldsymbol{\theta})\|q_\phi(\boldsymbol{d}|\boldsymbol{\theta})\right], \tag{2}$$

up to a constant. Consequently, when trained in this manner the density estimator learns to approximate the likelihood over the parameter space covered by the proposal distribution. For this work we use normalizing flows [56] as the conditional neural density estimator. Specifically, we adopt a masked autoregressive flow (MAF; [57]) constructed out of masked autoencoders for density estimation (MADE; [58]). In terms of implementation, we make use of the `sbi`[7] software package [59] to construct and train the NLE density estimator.

We follow the same simulation procedure as described in Section II A. Specifically, we make use of `CosmoPower-JAX` to simulate the matter power spectrum, `jax-cosmo` to compute the shear power spectrum, and a modified version of `sbi_lens` to then generate correlated log-normal convergence maps with Gaussian noise.

While modern neural density estimators can scale to relatively high dimensional settings, for field-level SBI it is typical to compress the field to a lower dimensional latent representation, for example by neural, statistical or wavelet scattering based compression techniques [*e.g.* 17, 19, 60–64]. While wavelet scattering transforms have recently been shown to be effective for this purpose and do not require additional simulations [21], for the purposes of this work we consider neural compression. An extensive study of neural compression techniques for SBI was recently performed by Ref. [7], demonstrating that a convolutional neural network (CNN) trained with a variational mutual information maximisation (VMIM; [17]) loss function can achieve excellent compression performance, capturing close to all higher order cosmological information in cosmic shear fields. We therefore adopt this neural compression technique and make use of the ResNet-18 CNN architecture [65] implemented in `Haiku`[8] [66] to compress the convergence maps, training our own compressor following the same procedure described in [67] and included in `sbi_lens`.

Finally, posterior samples are then generated by MCMC sampling using the surrogate likelihood. In this case, for simplicity we use the `emcee`[9] software package [68] to perform sampling, although alternative accelerated sampling techniques could be considered.

Our overall pipeline, including the simulator, is automatically differentiable, which can provide numerous advantages. While an automatically differentiable simulator is not strictly necessary for the SBI results presented in this work, it could in principle be used to train the NLE model more efficiently, requiring less training data [23, 24]. However, Ref. [67] recently found that incorporating gradients provided by automatic differentiation did not significantly improve field-level SBI inference and so we have not considered this further in the current article. Alternatively, the differentiable forward model is essential for field-level likelihood-based inference with a Bayesian hierarchical model (BHM) in order to leverage high-dimensional sampling techniques that exploit gradient information (*e.g.* NUTS). However, such an approach does not at present support Bayesian model selection and so we have not considered it further in the current article. We will present a field-level BHM approach that can also provide the calculation of the Bayesian evidence in an upcoming work.

C. Bayesian evidence for model comparison

For both settings considered previously, namely for both power spectrum likelihood-based inference and field-level SBI inference, we recover posterior samples by MCMC sampling. Moreover, for each sample the unnormalized posterior density will be evaluated during sampling. Thus, we have access to everything needed to compute the Bayesian evidence using the learned harmonic mean estimator, irrespective of the underlying method used to generate the posterior samples.

The Bayesian evidence is given by the marginalised likelihood

$$z = p(\boldsymbol{d}|M) = \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{d}|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M), \tag{3}$$

for likelihood $p(\boldsymbol{d}|\boldsymbol{\theta}, M)$ and prior $p(\boldsymbol{\theta}|M)$, where here we have made the model M explicit. The evidence is a critical term to compute in order to compare models. The posterior model odds between two competing models M_1 and M_2 can be written as

$$\frac{p(M_1|\boldsymbol{d})}{p(M_2|\boldsymbol{d})} = \frac{p(\boldsymbol{d}|M_1)p(M_1)}{p(\boldsymbol{d}|M_2)p(M_2)}, \tag{4}$$

which follows by Bayes' theorem. In many cases *a priori* probabilities $p(M_1)$ and $p(M_2)$ of the two models are considered to be equal, hence the ratio of posterior distributions becomes equivalent to the evidence ratio or Bayes factor

$$B_{12} = \frac{p(\boldsymbol{d}|M_1)}{p(\boldsymbol{d}|M_2)} = \frac{z_1}{z_2}. \tag{5}$$

For notational brevity, henceforth we drop the explicit conditioning on models unless there are multiple models under consideration.

The learned harmonic mean can be used to compute the evidence for different models, and thus to also compute Bayes factors for Bayesian model comparison. While the original harmonic mean [69] suffered from an exploding variance [70],

[7] `https://github.com/sbi-dev/sbi`
[8] `https://github.com/google-deepmind/dm-haiku`
[9] `https://github.com/dfm/emcee`

the learned harmonic mean solves this issue by integrating machine learning to learn an internal target distribution [35]. Critically, the learned internal target must be concentrated within the posterior. Normalizing flows provide an elegant way to ensure this simply by lowering the temperature T (*i.e.* variance) of their base distribution [37], avoiding the need for bespoke training. Given the learned target distribution $\varphi_{\psi}(\boldsymbol{\theta};T)$ with parameters ψ (i.e. neural network weights), the reciprocal evidence $\rho = z^{-1}$ is then estimated as

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N}\frac{\varphi_{\phi}(\theta_i;T)}{p(\boldsymbol{d}|\boldsymbol{\theta}_i,M)p(\boldsymbol{\theta}_i|M)}, \quad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}|\boldsymbol{d},M). \tag{6}$$

We compute evidence estimates from posterior samples for both likelihood-based and SBI settings using the `harmonic`[10] software package implementing the learned harmonic mean.

It is important to note that evidence values are of course sensitive to the choice of priors. This is a feature of Bayesian model comparison and not a bug, as it encapsulates Occam's razor [71]. In the Bayesian formalism models are specified as probability distributions over datasets and, since probability distributions must be normalized, each model has a limited "probability budget" to allocate. While a complex model can represent a wide range of datasets well, it spreads its predictive probability widely. In doing so, the model evidence of complex models will be penalised if such complexity is not required. There are a wide variety of ways to set priors appropriately for Bayesian inference depending on the statistical question at hand [72]. For example, approaches to setting priors include physical priors (*e.g.* non-negative mass or flux; [73]), uninformative Jeffreys priors that are invariant to a parameter transformation [74], informative priors for example to regularize inverse problems [*e.g.* 75], data-driven priors potentially specified by a generative model [*e.g.* 42, 76, 77], or data-informed priors, where the posterior of an *a priori* analysis is used as the prior for an analysis with new data [*e.g.* 78]. Despite a variety of methods to set appropriate piors, there remains debate regarding sensitivity of the evidence to prior choice [79–81]. If one wishes to remove the prior dependence for the purpose of studying tensions between data-sets, once the evidence is estimated it can be used to compute the Bayesian suspiciousness [82, 83].

III. RESULTS

We apply the power spectrum likelihood-based and field-level SBI inference frameworks outlined previously to simulated cosmic shear observatives intended to mimic a Stage IV survey. We run MCMC sampling to generate posterior samples, which we also use for evidence estimation with the learned harmonic mean estimator to compare wCDM and w_0w_aCDM models to ΛCDM. We present marginal posterior distributions of the cosmological parameters and Bayes factors for comparisons between the models considered, for a variety of ground truth data vectors.

While we follow the general methodology outlined in Section II, specific details of the simulator, neural compressor, neural density estimator, MCMC samplers, and learned harmonic mean estimator can be found in Appendix A.

A. Models, mock data & priors

We consider three models, the ubiquitous ΛCDM and wCDM cosmological models and also a phenomenological model that allows the equation of state for dark energy to evolve. For this dynamical dark energy we adopt the Chevallier-Polarski-Linder (CPL) parameterisation with $w(a) = w_0 + w_a(1-a)$ [84, 85], where a denotes the scale factor, resulting in the so-called w_0w_aCDM model. For model comparison, we focus on comparing wCDM and w_0w_aCDM models to ΛCDM. Recent results presented by the DESI collaboration [1] provide exciting hints of w_0w_aCDM, although this is only for certain data combinations and thus is far from conclusive.

For the two model comparisons performed (ΛCDM vs wCDM and ΛCDM vs w_0w_aCDM) we consider two different ground truth mock data cases, one generated by each model, resulting in four total model comparisons. Mock data cosmological parameter values and prior ranges are shown in Table I. In our analysis for all parameters besides (w_0, w_a) we consider the same priors as Ref. [86], matching the Rubin-LSST science requirements document. The ground truth is set to the middle of the prior range. We set $w = -1$ in the ΛCDM case. For other models we consider fiducial parameters as if the DESI results hinting at dynamical dark energy were the ground truth. That is, we set the (w_0, w_a) ground truth to the best-fit DESI parameters for the data combination showing hints of dynamical dark energy (DESI + CMB + PantheonPlus data) [1] for both the wCDM (with $w_0 = w$) and w_0w_aCDM cases. For (w_0, w_a), the posterior distributions of the Dark Energy Survey (DES; [87]) year three data [88] are used as the prior, *i.e.* we follow a data-informed prior approach (see Section II C).

B. ΛCDM vs wCDM

Figure 1 and Figure 2 show marginalised posterior distributions recovered for the power spectrum likelihood-based inference and field-level SBI inference, respectively, for ΛCDM vs wCDM. Both figures show results for the two different ground truth mock data cases. Bayes factors for each setting are displayed on each marginal distribution plot. Furthermore, they are summarised visually in Figure 3 for this ΛCDM vs wCDM comparison. Bayes factors for all experiments are summarised in Table II, complemented by the corresponding Jeffreys scale [89, 90] and odds ratio (assuming identical prior model probabilities).

It is apparent from the Bayes factors that it is not possible to distinguish between ΛCDM and wCDM models using the

[10] `https://github.com/astro-informatics/harmonic`

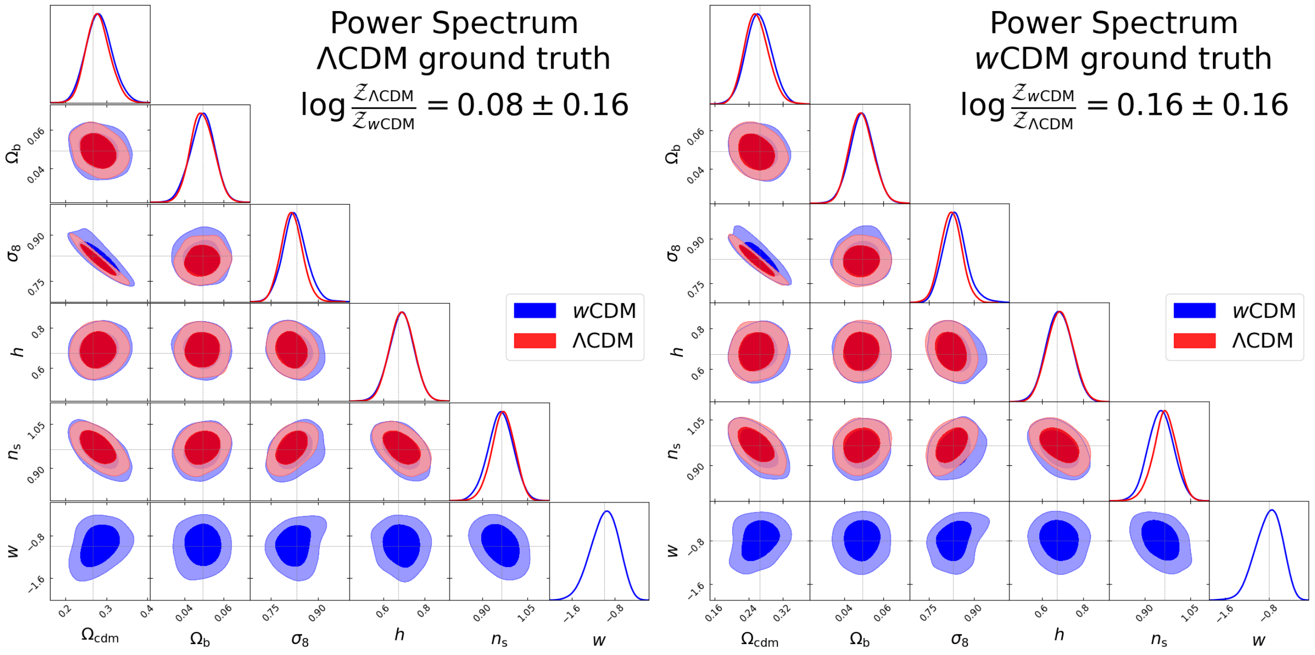

FIG. 1. Marginal posterior distributions of cosmological parameters for the **power spectrum likelihood-based inference**, comparing **ΛCDM vs *w*CDM**. Ground truth underlying parameter values are indicated by dashed lines. Left: ΛCDM ground truth data vector. Right: *w*CDM ground truth data vector. For both ground truth scenarios the Bayesian evidence values show it is **not possibile to distinguish cosmological models**.

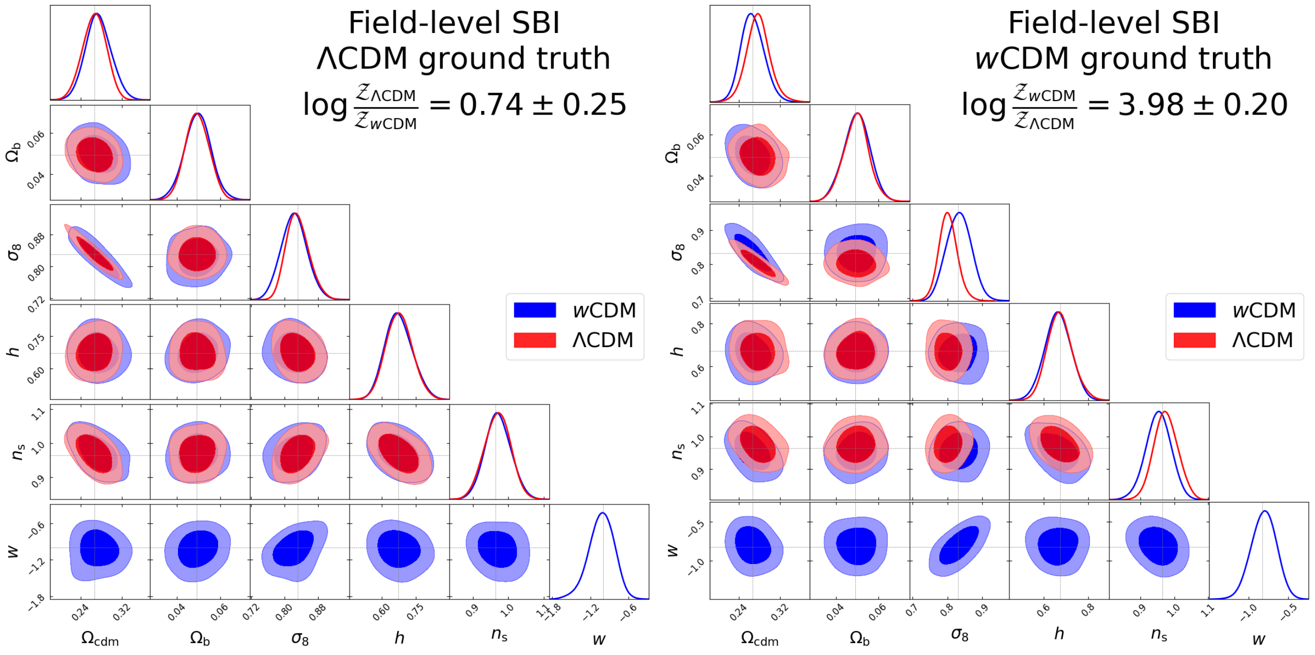

FIG. 2. Marginal posterior distributions of cosmological parameters for the **field-level SBI inference**, comparing **ΛCDM vs *w*CDM**. Ground truth underlying parameter values are indicated by dashed lines. Left: ΛCDM ground truth data vector. Right: *w*CDM ground truth data vector. For the former ground truth scenario the Bayesian evidence **weakly prefers the true underlying model ΛCDM**. For the latter ground truth scenario the Bayesian evidence **strongly pefers the true underlying model *w*CDM**.

TABLE I. Cosmological parameter values for the mock data and prior ranges. The normal distribution is denoted $\mathcal{N}$, while a truncated normal is denoted $\mathcal{N}_T$. The distribution for $\Omega_{\rm cdm}$ is truncated to have a lower bound of -1. The distribution for w is truncated to have a lower bound of -2.0 and an upper bound of -0.33.

Parameter	Mock data	Prior range
$\Omega_{\rm cdm}$	0.2664	$\mathcal{N}_T(0.2664, 0.2)$
$\Omega_{\rm b}$	0.0492	$\mathcal{N}(0.0492, 0.006)$
σ_8	0.831	$\mathcal{N}(0.831, 0.14)$
h	0.6727	$\mathcal{N}(0.6727, 0.063)$
$n_{\rm s}$	0.9645	$\mathcal{N}(0.9645, 0.08)$
w (ΛCDM)	-1.0	-
w (wCDM)	-0.827	$\mathcal{N}_T(-1.0, 0.9)$
w_0 (w_0w_aCDM)	-0.827	$\mathcal{N}(-0.95, 0.08)$
w_a (w_0w_aCDM)	-0.75	$\mathcal{N}(-0.4, 0.4)$

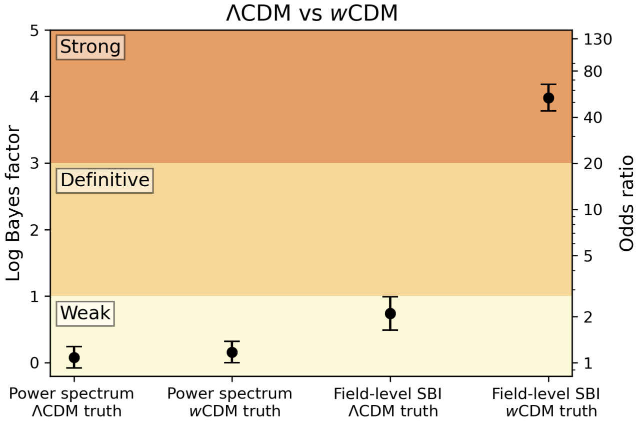

FIG. 3. Bayes factors with errors for the ΛCDM vs wCDM comparison. The shaded regions correspond to the strength of the Bayes factor on the Jeffreys scale. Note that **power spectrum likelihood-based inference cannot distinguish between ΛCDM and wCDM, whereas field-level SBI inference can**.

power spectrum alone, for either ΛCDM or wCDM ground truth mock data.

In contrast, for the field-level SBI inference it is possible to distinguish between ΛCDM and wCDM. Evidence for the correct underlying ground truth model is nevertheless considered weak on the Jeffreys scale at an odds ratio of 2.10:1 for the ΛCDM mock data, but it is strong with an odds ratio of 53.5:1 for the wCDM mock data. Furthermore, mismatches in data and model are now distinguishable visually from the contours of Figure 2, with the incorrect model showing a clear bias. Of course, when analysing real data the true underlying model is not known and so the evidence must be used for model comparison. The enhanced model constraining power of the field-level SBI analysis due to the extraction of high-order cosmological information is clear.

C. ΛCDM vs w_0w_aCDM

Figure 4 and Figure 5 show marginalised posterior distributions recovered for the power spectrum likelihood-based inference and field-level SBI inference, respectively, for ΛCDM vs w_0w_aCDM. Both figures show results for the two different ground truth mock data cases. Bayes factors for each setting are displayed on each marginal distribution plot. Furthermore, they are summarised visually in Figure 6 and also included in Table II.

Similar to when comparing ΛCDM and wCDM, it is apparent from the Bayes factors that it is not possible to distinguish between ΛCDM and w_0w_aCDM models using the power spectrum alone, for either ΛCDM or w_0w_aCDM ground truth mock data.

In contrast, for the field-level SBI inference it is possible to distinguish between ΛCDM and w_0w_aCDM and the correct underlying model is selected. On the Jeffreys scale model selection is definitive for both ground truth mock data scenarios. It should also be noted that posterior contours are also more accurately centred on the ground truth for the field-level SBI inference, unlike for the power spectrum inference.

IV. CONCLUSIONS

We present a framework that for the first time allows Bayesian model comparison to be performed for field-level inference of cosmological models. We achieve this by leveraging SBI so that a reduced parameter space containing only cosmological parameters of interest need be considered. This reduces the dimensionality of the parameter space considerably compared to likelihood-based field-level inference where the pixels of the initial or observed field are treated as parameters to be inferred. Specifically, we take an NLE (neural likelihood estimation) approach, training a density estimator to learn a surrogate for the likelihood, using the `CosmoPower` emulator [22, 47] to accelerate the generation of simulations needed for training. We then perform MCMC sampling to generate posterior samples, which are not only used for parameter estimator but to also compute the Bayesian evidence for model selection using the learned harmonic mean estimator implemented in the `harmonic` code [35, 37, 43].

We apply our framework to mock Stage IV cosmic shear observations to assess its effectiveness at distinguishing between various models of dark energy. For comparison purposes we also consider a traditional power spectrum likelihood-based inference. Due to the higher order statistical information extracted by our field-level SBI approach, it is able to successfully distinguish dynamical dark energy from ΛCDM using the Bayesian evidence, whereas the power spectrum inference cannot. If the DESI results that hinted at the possibility of dynamical dark energy [1] were indeed the true underlying model, Stage IV surveys such as those by Euclid and Rubin-LSST, would be able to provide definitive evidence for dynamical dark energy.

Given the effectiveness of our field-level cosmological model selection framework, it is important to extend it to a

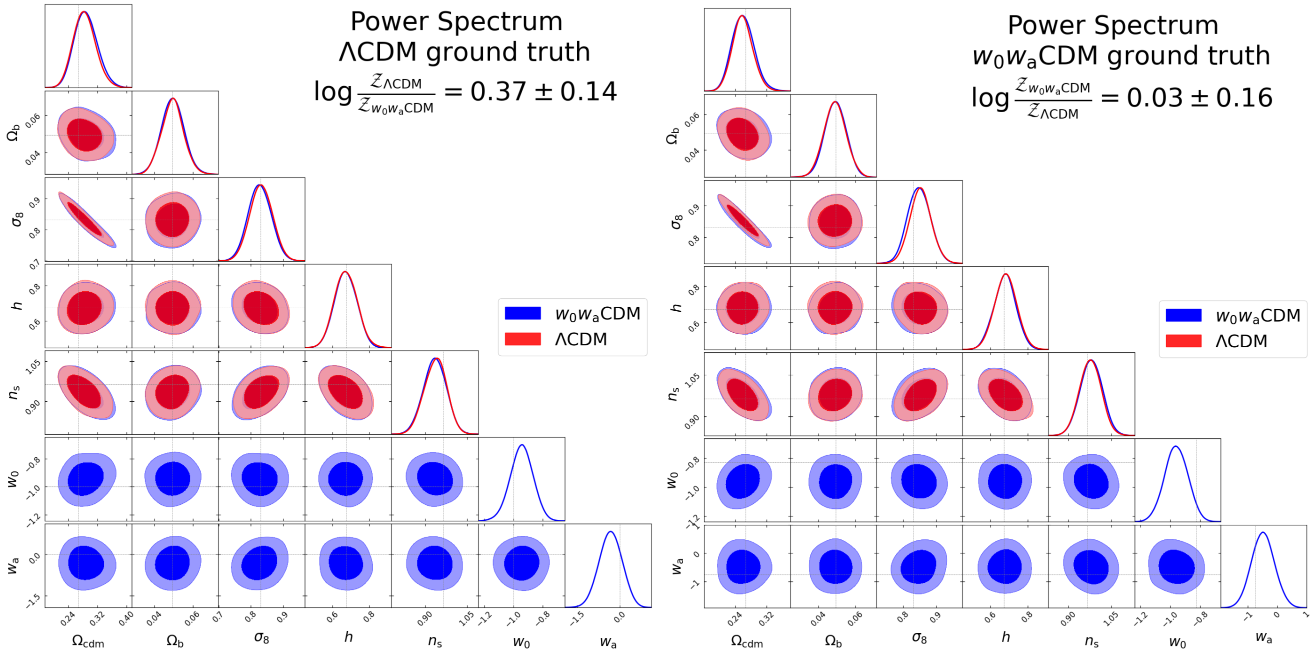

FIG. 4. Marginal posterior distributions of cosmological parameters for the **power spectrum likelihood-based inference**, comparing **ΛCDM vs** w_0w_a**CDM**. Ground truth underlying parameter values are indicated by dashed lines. Left: ΛCDM ground truth data vector. Right: w_0w_aCDM ground truth data vector. For both ground truth scenarios the Bayesian evidence values show it is **not possibile to distinguish cosmological models**.

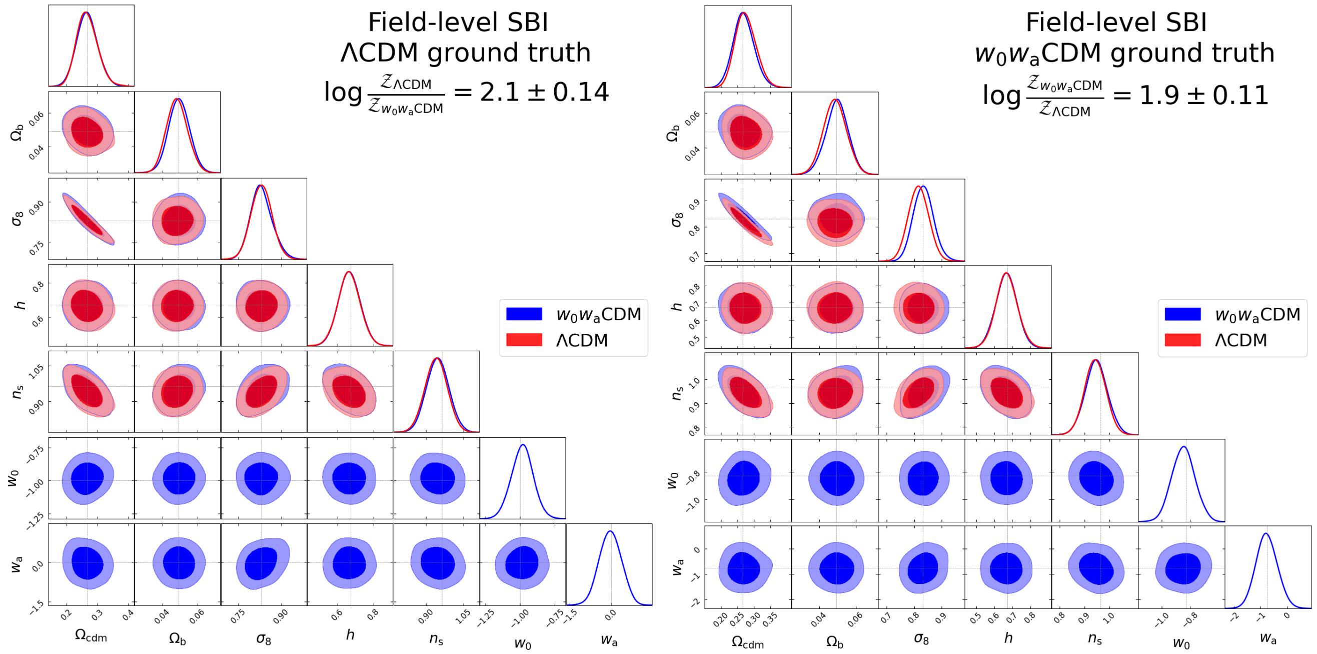

FIG. 5. Marginal posterior distributions of cosmological parameters for the **field-level SBI inference**, comparing **ΛCDM vs** w_0w_a**CDM**. Ground truth underlying parameter values are indicated by dashed lines. Left: ΛCDM ground truth data vector. Right: w_0w_aCDM ground truth data vector. For the former ground truth scenario the Bayesian evidence **definitively prefers the true underlying model ΛCDM**. For the latter ground truth scenario the Bayesian evidence **definitively pefers the true underlying model** w_0w_a**CDM**.

TABLE II. Summary of Bayes factors, where they lie on the Jeffreys scale and corresponding odds ratio for the different model comparisons and ground truth data vector model combinations considered. Notably we can see that using **power spectrum likelihood-based inference cannot distinguish between the cosmological models**, whereas **field-level SBI inference can distinguish between cosmological models**.

Model	Ground truth	Method	Bayes factor (log)	Jeffreys scale	Odds ratio
ΛCDM vs wCDM	ΛCDM	Power spectrum	0.08 ± 0.16	inconclusive	1.08 : 1
ΛCDM vs wCDM	wCDM	Power spectrum	0.16 ± 0.16	inconclusive	1.17 : 1
ΛCDM vs wCDM	ΛCDM	Field-level SBI	0.74 ± 0.25	weak	2.10 : 1
ΛCDM vs wCDM	wCDM	Field-level SBI	3.98 ± 0.20	strong	53.5 : 1
ΛCDM vs w_0w_aCDM	ΛCDM	Power spectrum	0.37 ± 0.14	inconclusive	1.45 : 1
ΛCDM vs w_0w_aCDM	w_0w_aCDM	Power spectrum	0.03 ± 0.16	inconclusive	1.03 : 1
ΛCDM vs w_0w_aCDM	ΛCDM	Field-level SBI	2.10 ± 0.14	definitive	8.17 : 1
ΛCDM vs w_0w_aCDM	w_0w_aCDM	Field-level SBI	1.90 ± 0.11	definitive	6.69 : 1

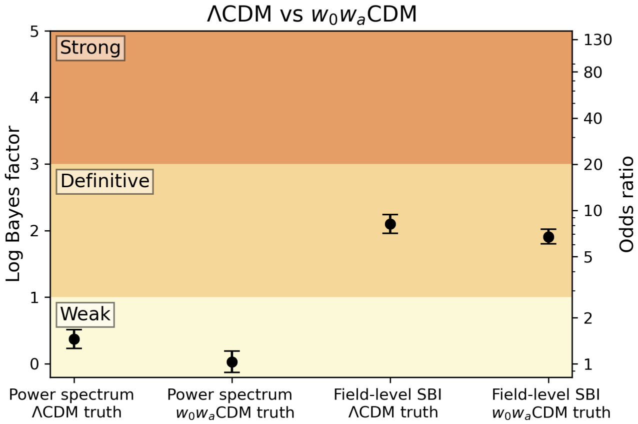

FIG. 6. Bayes factors with errors for the ΛCDM vs w_0w_aCDM comparison. The shaded regions correspond to the strength of the Bayes factor on the Jeffreys scale. Note that **power spectrum likelihood-based inference cannot distinguish between ΛCDM and w_0w_aCDM, whereas field-level SBI inference can**.

more realistic setting in preparation for application to Stage IV surveys. In particular, more realistic simulations, observational effects and systematics need to be incorporated in the forward model. Furthermore, the forward modelling, any field-level emulation, and compression must be extended to the spherical setting to support the wide fields of upcoming surveys (*e.g.* using spherical machine learning or scattering techniques; [91–94]). Field-level SBI techniques exhibit significant promise and when applied to upcoming Stage IV surveys could provide one of the most effective means of determining the underlying nature of dark energy.

ACKNOWLEDGMENTS

KL is supported by STFC (grant number ST/W001136/1). JDM is supported by EPSRC (grant number EP/W007673/1) and STFC (grant number ST/W001136/1).

[1] A. Adame, J. Aguilar, S. Ahlen, S. Alam, D. Alexander, M. Alvarez, O. Alves, A. Anand, U. Andrade, E. Armengaud, *et al.*, Desi 2024 vi: Cosmological constraints from the measurements of baryon acoustic oscillations, arXiv preprint arXiv:2404.03002 (2024).

[2] S. S. Boruah, E. Rozo, and P. Fiedorowicz, Map-based cosmology inference with lognormal cosmic shear maps, Monthly Notices of the Royal Astronomical Society **516**, 4111 (2022).

[3] E. Tsaprazi, N.-M. Nguyen, J. Jasche, F. Schmidt, and G. Lavaux, Field-level inference of galaxy intrinsic alignment from the sdss-iii boss survey, Journal of Cosmology and Astroparticle Physics **2022** (08), 003.

[4] N. Porqueres, A. Heavens, D. Mortlock, and G. Lavaux, Lifting weak lensing degeneracies with a field-based likelihood, Monthly Notices of the Royal Astronomical Society **509**, 3194 (2022).

[5] A. Andrews, J. Jasche, G. Lavaux, and F. Schmidt, Bayesian field-level inference of primordial non-gaussianity using next-generation galaxy surveys, Monthly Notices of the Royal Astronomical Society **520**, 5746 (2023).

[6] N. Porqueres, A. Heavens, D. Mortlock, G. Lavaux, and T. L. Makinen, Field-level inference of cosmic shear with intrinsic alignments and baryons, arXiv preprint arXiv:2304.04785 (2023).

[7] D. Lanzieri, J. Zeghal, T. L. Makinen, A. Boucaud, J.-L. Starck, and F. Lanusse, Optimal neural summarisation for full-field weak lensing cosmological implicit inference, arXiv preprint arXiv:2407.10877 (2024).

[8] N.-M. Nguyen, F. Schmidt, B. Tucci, M. Reinecke, and A. Kostić, How much information can be extracted from galaxy clustering at the field level?, arXiv e-prints , arXiv:2403.03220 (2024), arXiv:2403.03220 [astro-ph.CO].

[9] R. Laureijs, J. Amiaux, S. Arduini, J.-L. Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, *et al.*, Euclid Definition Study Report, arXiv e-prints , arXiv:1110.3193 (2011), arXiv:1110.3193 [astro-ph.CO].

[10] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. Asztalos, T. S. Axelrod, S. Bailey, *et al.*, LSST Science Book, Version 2.0, arXiv e-prints , arXiv:0912.0201 (2009), arXiv:0912.0201 [astro-ph.IM].

[11] D. Spergel, N. Gehrels, J. Breckinridge, M. Donahue,

A. Dressler, B. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, *et al.*, Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report, arXiv e-prints , arXiv:1503.03757 (2015), arXiv:1503.03757 [astro-ph.IM].

[12] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proceedings of the National Academy of Sciences **117**, 30055 (2020).

[13] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, Fast likelihood-free cosmology with neural density estimators and active learning, Monthly Notices of the Royal Astronomical Society **488**, 4440 (2019).

[14] P. L. Taylor, T. D. Kitching, J. Alsing, B. D. Wandelt, S. M. Feeney, and J. D. McEwen, Cosmic shear: Inference from forward models, Physical Review D **100**, 023519 (2019).

[15] K. Lin, M. von Wietersheim-Kramsta, B. Joachimi, and S. Feeney, A simulation-based inference pipeline for cosmic shear with the Kilo-Degree Survey, Monthly Notices of the Royal Astronomical Society **524**, 6167 (2023).

[16] M. von Wietersheim-Kramsta, K. Lin, N. Tessore, B. Joachimi, A. Loureiro, R. Reischke, and A. H. Wright, KiDS-SBI: Simulation-based inference analysis of kids-1000 cosmic shear, ArXiv (2024).

[17] N. Jeffrey, J. Alsing, and F. Lanusse, Likelihood-free inference with neural compression of des sv weak lensing map statistics, Monthly Notices of the Royal Astronomical Society **501**, 954 (2021).

[18] P. Lemos, L. Parker, C. Hahn, S. Ho, M. Eickenberg, J. Hou, E. Massara, C. Modi, A. Moradinezhad Dizgah, B. Régaldo-Saint Blancard, *et al.*, Simbig: Field-level simulation-based inference of large-scale structure, Machine Learning for Astrophysics , 18 (2023).

[19] M. Gatti, N. Jeffrey, L. Whiteway, J. Williamson, B. Jain, V. Ajani, D. Anbajagane, G. Giannini, C. Zhou, A. Porredon, *et al.*, Dark energy survey year 3 results: Simulation-based cosmological inference with wavelet harmonics, scattering transforms, and moments of weak lensing mass maps. validation on simulations, Physical Review D **109**, 063534 (2024).

[20] N. Jeffrey, L. Whiteway, M. Gatti, J. Williamson, J. Alsing, A. Porredon, J. Prat, C. Doux, B. Jain, C. Chang, *et al.*, Dark energy survey year 3 results: likelihood-free, simulation-based *w* cdm inference with neural compression of weak-lensing map statistics, arXiv preprint arXiv:2403.02314 (2024).

[21] K. Lin, B. Joachimi, and J. D. McEwen, Simulation-based inference with scattering representations: scattering is all you need, in *Proceedings of the Machine Learning and Physical Sciences Workshop as part of the 38th International Conference on Neural Information Processing Systems* (2024).

[22] A. Spurio Mancini, D. Piras, J. Alsing, B. Joachimi, and M. P. Hobson, Cosmopower: emulating cosmological power spectra for accelerated bayesian inference from next-generation surveys, Monthly Notices of the Royal Astronomical Society **511**, 1771–1788 (2022).

[23] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, Mining gold from implicit models to improve likelihood-free inference, Proceedings of the National Academy of Sciences **117**, 5242 (2020).

[24] J. Zeghal, F. Lanusse, A. Boucaud, B. Remy, and E. Aubourg, Neural posterior estimation with differentiable simulators, arXiv preprint arXiv:2207.05636 (2022).

[25] R. Trotta, Applications of Bayesian model selection to cosmological parameters, Monthly Notices of the Royal Astronomical Society **378**, 72 (2007), https://academic.oup.com/mnras/article-pdf/378/1/72/3961005/mnras0378-0072.pdf.

[26] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Analysis **1**, 833 (2006).

[27] F. Feroz and M. P. Hobson, Multimodal nested sampli:w ng: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses, Monthly Notices of the Royal Astronomical Society **384**, 449 (2008), https://academic.oup.com/mnras/article-pdf/384/2/449/3378518/mnras0384-0449.pdf.

[28] W. J. Handley, M. P. Hobson, and A. N. Lasenby, polychord: nested sampling for cosmology, Monthly Notices of the Royal Astronomical Society: Letters **450**, L61 (2015), https://academic.oup.com/mnrasl/article-pdf/450/1/L61/3087909/slv047.pdf.

[29] W. J. Handley, M. P. Hobson, and A. N. Lasenby, polychord: next-generation nested sampling, Monthly Notices of the Royal Astronomical Society **453**, 4384 (2015), https://academic.oup.com/mnras/article-pdf/453/4/4384/8034904/stv1911.pdf.

[30] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, Importance nested sampling and the MultiNest algorithm, The Open Journal of Astrophysics **2**, 10.21105/astro.1306.2144 (2019).

[31] E. Higson, W. Handley, M. Hobson, and A. Lasenby, Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation, Statistics and Computing **29**, 891 (2019).

[32] J. S. Speagle, dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences, Monthly Notices of the Royal Astronomical Society **493**, 3132 (2020).

[33] X. Cai, J. D. McEwen, and M. Pereyra, Proximal nested sampling for high-dimensional bayesian model selection, Statistics and Computing **32**, 87 (2022).

[34] J. U. Lange, nautilus: boosting Bayesian importance nested sampling with deep learning, Monthly Notices of the Royal Astronomical Society **525**, 3181 (2023).

[35] J. D. McEwen, C. G. Wallis, M. A. Price, and A. S. Mancini, Machine learning assisted bayesian model comparison: learnt harmonic mean estimator, arXiv preprint arXiv:2111.12720 (2021).

[36] A. Polanska, M. A. Price, A. Spurio Mancini, and J. D. McEwen, Learned harmonic mean estimation of the marginal likelihood with normalizing flows, Physical Sciences Forum **9**, 10.3390/psf2023009010 (2023).

[37] A. Polanska, M. A. Price, D. Piras, A. S. Mancini, and J. D. McEwen, Learned harmonic mean estimation of the bayesian evidence with normalizing flows (2024), arXiv:2405.05969 [astro-ph.IM].

[38] D. Piras, A. Polanska, A. S. Mancini, M. A. Price, and J. D. McEwen, The future of cosmological likelihood-based inference: accelerated high-dimensional parameter estimation and model comparison, arXiv preprint arXiv:2405.12965 (2024).

[39] M. D. Hoffman and A. Gelman, The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, J. Mach. Learn. Res. **15**, 1593–1623 (2014).

[40] K. W. k. Wong, M. Gabrié, and D. Foreman-Mackey, flowMC: Normalizing flow enhanced sampling package for probabilistic inference in JAX, J. Open Source Softw. **8**, 5021 (2023), arXiv:2211.06397 [astro-ph.IM].

[41] A. Polanska, T. Wouters, P. T. H. Pang, K. W. K. Wong, and J. D. McEwen, Accelerated bayesian parameter estimation and model selection for gravitational waves with normalizing flows, in *Proceedings of the Machine Learning and Physical Sciences Workshop as part of the 38th International Conference on Neural Information Processing Systems* (2024).

[42] J. D. McEwen, T. I. Liaudat, M. A. Price, X. Cai, and M. Pereyra,

Proximal nested sampling with data-driven priors for physical scientists, in *Physical Sciences Forum*, Vol. 9 (MDPI, 2023) p. 13.

[43] A. Spurio Mancini, M. Docherty, M. Price, and J. McEwen, Bayesian model comparison for simulation-based inference, RAS Techniques and Instruments **2**, 710 (2023).

[44] N. Jeffrey and B. D. Wandelt, Evidence networks: simple losses for fast, amortized, neural bayesian model comparison, Machine Learning: Science and Technology **5**, 015008 (2024).

[45] S. T. Radev, M. D'Alessandro, U. K. Mertens, A. Voss, U. Köthe, and P.-C. Bürkner, Amortized bayesian model comparison with evidential deep learning, IEEE Transactions on Neural Networks and Learning Systems **34**, 4903 (2021).

[46] G. Papamakarios, D. Sterratt, and I. Murray, Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, in *The 22nd international conference on artificial intelligence and statistics* (PMLR, 2019) pp. 837–848.

[47] D. Piras and A. Spurio Mancini, CosmoPower-JAX: high-dimensional Bayesian inference with differentiable cosmological emulators, The Open Journal of Astrophysics **6** (2023).

[48] J.-E. Campagne, F. Lanusse, J. Zuntz, A. Boucaud, S. Casas, M. Karamanis, D. Kirkby, D. Lanzieri, Y. Li, and A. Peel, Jax-cosmo: An end-to-end differentiable and gpu accelerated cosmology library, arXiv preprint arXiv:2302.05163 (2023).

[49] A. Mead, C. Heymans, L. Lombriser, J. Peacock, O. Steele, and H. Winther, Accurate halo-model matter power spectra with dark energy, massive neutrinos and modified gravitational forces, Monthly Notices of the Royal Astronomical Society **459**, 1468 (2016).

[50] A. Petri, Mocking the weak lensing universe: The LensTools Python computing package, Astronomy and Computing **17**, 73 (2016), arXiv:1606.01903 [astro-ph.CO].

[51] D. Phan, N. Pradhan, and M. Jankowiak, Composable effects for flexible and accelerated probabilistic programming in numpyro, arXiv preprint arXiv:1912.11554 (2019).

[52] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, Pyro: Deep universal probabilistic programming, J. Mach. Learn. Res. **20**, 28:1 (2019).

[53] G. Papamakarios and I. Murray, Fast ϵ-free inference of simulation models with bayesian conditional density estimation, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16 (Curran Associates Inc., Red Hook, NY, USA, 2016) p. 1036–1044.

[54] J.-M. Lueckmann, P. J. Gonçalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, Flexible statistical inference for mechanistic models of neural dynamics, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (Curran Associates Inc., Red Hook, NY, USA, 2017) p. 1289–1299.

[55] D. Greenberg, M. Nonnenmacher, and J. Macke, Automatic posterior transformation for likelihood-free inference, in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 97, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, 2019) pp. 2404–2414.

[56] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, The Journal of Machine Learning Research **22**, 2617 (2021).

[57] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, Advances in neural information processing systems **30** (2017).

[58] M. Germain, K. Gregor, I. Murray, and H. Larochelle, Made: Masked autoencoder for distribution estimation, in *International conference on machine learning* (PMLR, 2015) pp. 881–889.

[59] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke, sbi: A toolkit for simulation-based inference, Journal of Open Source Software **5**, 2505 (2020).

[60] M. Tegmark, A. N. Taylor, and A. F. Heavens, Karhunen-Loeve eigenvalue problems in cosmology: How should we tackle large data sets?, The Astrophysical Journal **480**, 22 (1997).

[61] A. F. Heavens, R. Jimenez, and O. Lahav, Massive lossless data compression and multiple parameter estimation from galaxy spectra, Monthly Notices of the Royal Astronomical Society **317**, 965 (2000).

[62] J. Alsing and B. Wandelt, Generalized massive optimal data compression, Monthly Notices of the Royal Astronomical Society: Letters **476**, L60 (2018).

[63] T. Charnock, G. Lavaux, and B. D. Wandelt, IMNN: Information maximizing neural networks, Astrophysics Source Code Library , ascl (2018).

[64] T. L. Makinen, J. Alsing, and B. D. Wandelt, Fishnets: Information-optimal, scalable aggregation for sets and graphs, arXiv preprint arXiv:2310.03812 (2023).

[65] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.

[66] T. Hennigan, T. Cai, T. Norman, L. Martens, and I. Babuschkin, Haiku: Sonnet for JAX (2020).

[67] J. Zeghal, D. Lanzieri, F. Lanusse, A. Boucaud, G. Louppe, E. Aubourg, A. E. Bayer, and T. L. Collaboration, Simulation-based inference benchmark for lsst weak lensing cosmology, arXiv preprint arXiv:2409.17975 (2024).

[68] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: the mcmc hammer, Publications of the Astronomical Society of the Pacific **125**, 306 (2013).

[69] M. A. Newton and A. E. Raftery, Approximate Bayesian inference with the weighted likelihood bootstrap, Journal of the Royal Statistical Society: Series B (Methodological) **56**, 3 (1994).

[70] R. M. Neal, Contribution to the discussion of "Approximate Bayesian inference with the weighted likelihood bootstrap" by Newton MA, Raftery AE, JR Stat Soc Ser A (Methodological) **56**, 41 (1994).

[71] Z. Ghahramani, Bayesian non-parametrics and the probabilistic approach to modelling, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**, 20110553 (2013).

[72] F. Llorente, L. Martino, E. Curbelo, J. Ló pez Santiago, and D. Delgado, On the safe use of prior densities for bayesian model selection, Wiley Interdisciplinary Reviews: Computational Statistics **15**, e1595 (2023).

[73] R. E. Carrillo, J. D. McEwen, and Y. Wiaux, Purify: a new approach to radio-interferometric imaging, Monthly Notices of the Royal Astronomical Society **439**, 3591 (2014).

[74] P. M. Lee, *Bayesian statistics* (Oxford University Press London:, 1989).

[75] M. A. Price, J. D. McEwen, L. Pratley, and T. D. Kitching, Sparse bayesian mass-mapping with uncertainties: Full sky observations on the celestial sphere, Monthly Notices of the Royal Astronomical Society **500**, 5436 (2021).

[76] B. Remy, F. Lanusse, N. Jeffrey, J. Liu, J.-L. Starck, K. Osato, and T. Schrabback, Probabilistic mass-mapping with neural score estimation, Astronomy & Astrophysics **672**, A51 (2023).

[77] T. I. Liaudat, M. Mars, M. A. Price, M. Pereyra, M. M. Betcke, and J. D. McEwen, Scalable bayesian uncertainty quantification with data-driven priors for radio interferometric imaging, RAS

Techniques and Instruments **3**, 505 (2024).

[78] J. Alsing and W. Handley, Nested sampling with any prior you like, Monthly Notices of the Royal Astronomical Society: Letters **505**, L95 (2021).

[79] E. V. Linder and R. Miquel, Tainted evidence: cosmological model selection vs. fitting, arXiv preprint astro-ph/0702542 (2007).

[80] A. R. Liddle, P. S. Corasaniti, M. Kunz, P. Mukherjee, D. Parkinson, and R. Trotta, Comment ontainted evidence: cosmological model selection versus fitting', by eric v. linder and ramon miquel (astro-ph/0702542v2), arXiv preprint astro-ph/0703285 (2007).

[81] G. Efstathiou, Limitations of bayesian evidence applied to cosmology, Monthly Notices of the Royal Astronomical Society **388**, 1314 (2008).

[82] W. Handley and P. Lemos, Quantifying tensions in cosmological parameters: Interpreting the des evidence ratio, Physical Review D **100**, 043504 (2019).

[83] P. Lemos, F. Köhlinger, W. Handley, B. Joachimi, L. Whiteway, and O. Lahav, Quantifying suspiciousness within correlated data sets, Monthly Notices of the Royal Astronomical Society **496**, 4647 (2020).

[84] M. Chevallier and D. Polarski, Accelerating Universes with Scaling Dark Matter, International Journal of Modern Physics D **10**, 213 (2001), arXiv:gr-qc/0009008 [gr-qc].

[85] E. V. Linder, Exploring the Expansion History of the Universe, Phys. Rev. Lett. **90**, 091301 (2003), arXiv:astro-ph/0208512 [astro-ph].

[86] Z. Zhang, C. Chang, P. Larsen, L. F. Secco, J. Zuntz, and L. D. E. S. Collaboration, Transitioning from stage-iii to stage-iv: cosmology from galaxy× cmb lensing and shear× cmb lensing, Monthly Notices of the Royal Astronomical Society **514**, 2181 (2022).

[87] D. E. S. C. Fermilab, U. of Illinois at Urbana-Champaign, U. of Chicago, L. B. N. Laboratory, C.-T. I.-A. Observatory, and B. Flaugher, The dark energy survey, International Journal of Modern Physics A **20**, 3121 (2005).

[88] T. Abbott, M. Aguena, A. Alarcon, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, S. Avila, D. Bacon, E. Baxter, *et al.*, Dark energy survey year 3 results: Constraints on extensions to λ cdm with weak lensing and galaxy clustering, Physical Review D **107**, 10.1103/physrevd.107.083504 (2023).

[89] H. Jeffreys, *The theory of probability* (OuP Oxford, 1998).

[90] S. Nesseris and J. Garcia-Bellido, Is the jeffreys' scale a reliable tool for bayesian model comparison in cosmology?, Journal of Cosmology and Astroparticle Physics **2013** (08), 036.

[91] O. J. Cobb, C. G. R. Wallis, A. N. Mavor-Parker, A. Marignier, M. Price, M. d'Avezac, and J. D. McEwen, Efficient generalized spherical CNNs, in *International Conference on Learning Representations (ICLR)* (2021) arXiv:2010.11661.

[92] J. Ocampo, M. A. Price, and J. D. McEwen, Scalable and equivariant spherical CNNs by discrete-continuous (DISCO) convolutions, in *International Conference on Learning Representations (ICLR)* (2023) arXiv:2209.13603.

[93] J. D. McEwen, C. G. R. Wallis, and A. N. Mavor-Parker, Scattering networks on the sphere for scalable and rotationally equivariant spherical CNNs, in *International Conference on Learning Representations (ICLR)* (2022) arXiv:2102.02828.

[94] L. Mousset, E. Allys, M. A. Price, J. Aumont, J.-M. Delouis, L. Montier, and J. D. McEwen, Generative models of astrophysical fields with scattering transforms on the sphere, Astronomy & Astrophysics, in press (2024), arXiv:2407.07007.

[95] The LSST Dark Energy Science Collaboration, R. Mandelbaum, T. Eifler, R. Hložek, T. Collett, E. Gawiser, D. Scolnic, D. Alonso, H. Awan, R. Biswas, J. Blazek, *et al.*, The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document, arXiv e-prints , arXiv:1809.01669 (2018), arXiv:1809.01669 [astro-ph.CO].

[96] I. Smail, D. W. Hogg, L. Yan, and J. G. Cohen, Deep Optical Galaxy Counts with the Keck Telescope, The Astrophysical Journal Letters **449**, L105 (1995), arXiv:astro-ph/9506095 [astro-ph].

[97] A. Lewis, Getdist: a python package for analysing monte carlo samples, arXiv preprint arXiv:1910.13970 (2019).

[98] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, Advances in neural information processing systems **32** (2019).

Appendix A: Configuration

While the general methodology is described in Section II, we outline here the specific settings and parameters configured to produce the results presented in Section III.

1. Survey settings

To simulate mock Stage IV survey data we follow the Rubin-LSST science requirements document [95] and target a Y10 data release. This means that the underlying source galaxy redshift distribution follows a Smail distribution [96] parameterised by

$$n(z) \propto z^2 \exp\left(-\frac{z}{z_0}\right)^{\alpha}, \tag{A1}$$

with $z_0 = 0.11$ and $\alpha = 0.68$ and with 5 redshift bins each containing an equal number of galaxies and photometric redshift error given by $\sigma_z = 0.05(1+z)$. To model survey observational noise, we assume a shape noise of $\sigma_e = 0.26$ and a galaxy number density of $n_g = 27$ arcmin^{-2}. Following Ref. [7] for `sbi_lens` we set the pixel area $A_{\text{pix}} = 5.49$ arcmin^2 and observed area to 10×10 deg^2. As such, we model survey noise as additive Gaussian noise with zero mean and variance per tomographic bin given by

$$\sigma^2_{\text{noise}} = \frac{\sigma_e^2}{n_g A_{\text{pix}}}. \tag{A2}$$

2. Compression

Compression is performed with a convolutional neural network with a ResNet-18 architecture [65]. The network is implemented in `Haiku` [66]. Following Ref. [7], we make use of a Variational Mutual Information Maximization (VMIM) loss function introduced to cosmology in Ref. [17] which is shown to produce sufficient statistics for SBI. We train our own compression with the aforementioned architecture following the same procedure described in [67] and included in `sbi_lens`.

3. SBI NLE density estimator

We make use of a masked autoregressive flow (MAF) [57] as the conditional neural density estimator. The MAF is constructed out of 5 masked autoencoders for density estimation (MADE) [58] with 50 hidden features each. The NLE density estimator is trained using the `sbi` software package [59]. Following the work of Ref. [7] we make use of 150,000 compressed simulations for training, which is likely more than strictly necessary.

4. MCMC

To obtain posterior samples, we make use of NUTS [39] implemented in `NumPyro` for the power spectrum analysis as detailed in Section II A and `emcee` [68] for the field-level SBI analysis as detailed in Section II B. For NUTS we set the number of chains to 3, with a burn-in length of 1200 and chain length of 1800. For `emcee` we run 24 walkers with 200 burn-in steps and 300 samples per walker. In both cases this results in 7200 samples after burn-in. We plot our contours with the `getdist`[11] software package [97].

5. Learned harmonic mean estimator

For the internal learned target distribution of the learned harmonic mean we train a rational quadratic spline flow [98], including standardization [37], consisting of 2 layers, with 128 spline bins. For the results presented in this article we concentrated the flow using a temperature of $T = 0.8$, although we also found overall results were robust to changing the temperature from 0.4 to 0.9 in steps of 0.1. Of the available MCMC samples, 50% were used for training the flow and 50% for evidence calculation.

[11] `https://github.com/cmbant/getdist`