How to embed any likelihood into SBI: Application to *Planck* + Stage IV galaxy surveys and Dynamical Dark Energy

Guillermo Franco Abellán,[1,*] Noemi Anau Montel,[2] Oleg Savchenko,[1] and Christoph Weniger[1]

[1] *GRAPPA Institute, Institute for Theoretical Physics Amsterdam,*
University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[2] *Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany*

(Dated: August 1, 2025)

Simulation-based inference (SBI) allows fast Bayesian inference for simulators encoding implicit likelihoods. However, some explicit likelihoods cannot be easily reformulated as simulators, hindering their integration into combined analyses within SBI frameworks. One key example in cosmology is given by the *Planck* CMB likelihoods. We present a simple method to construct an effective simulator for any explicit likelihood using samples from a previously converged Markov Chain Monte Carlo (MCMC) run. This effective simulator can subsequently be combined with any forward simulator. To illustrate this method, we combine the full *Planck* CMB likelihoods with a 3×2pt simulator (cosmic shear, galaxy clustering and their cross-correlation) for a Stage IV survey like *Euclid*, and test evolving dark energy parameterized by the w_0w_a equation-of-state. Assuming the w_0w_aCDM cosmology hinted by DESI BAO DR2 + *Planck* 2018 + PantheonPlus SNIa datasets, we find that future 3×2pt data alone could detect evolving dark energy at 5σ, while its combination with current CMB, BAO and SNIa datasets could raise the detection to almost 7σ. Moreover, thanks to simulation reuse enabled by SBI, we show that our joint analysis is in excellent agreement with MCMC while requiring *zero* Boltzmann solver calls. This result opens up the possibility of performing massive global scans combining explicit and implicit likelihoods in a highly efficient way.

I. INTRODUCTION

The most stringent and robust cosmological results of the next decade will be obtained from the combination of multiple astronomical surveys. This includes data from past and current experiments like *Planck* [1], ACT [2], SPT [3], KiDS [4], DES [5], BOSS [6], and DESI [7], as well as forthcoming surveys such as *Euclid* [8], LSST [9], Simons Observatory [10], and CMB-S4 [11]. Performing Bayesian inference from these surveys typically involves evaluating a large number of theoretical models – each requiring costly calls to Boltzmann solvers like `CAMB` [12] or `CLASS` [13] – and marginalizing over numerous nuisance parameters. In the context of Markov Chain Monte Carlo (MCMC) methods, this time consuming exercise needs to be repeated for each new combination of experiments,[1] which becomes increasingly challenging given the diversity and complexity of datasets.

In recent years, simulation-based inference (SBI) [16] has emerged as a powerful cosmological tool for Bayesian analysis in complex settings (see e.g. [17–25]). By using forward simulations that encode implicit likelihoods, SBI offers several key benefits. First, it supports amortized inference for rapid re-analysis and cross-validation. Second, it can flexibly incorporate complex effects through forward modeling that would be expensive to compute or intractable to express analytically in a likelihood. Third, it can efficiently yield marginal parameter estimates by directly integrating out nuisance variables. Finally, because simulations can be reused across inference tasks, SBI eliminates the need for repeated model evaluations, greatly enhancing computational efficiency.

However, some legacy cosmological likelihoods, even if explicit in form, cannot be readily reformulated as forward simulators. A notable example are the *Planck* cosmic microwave background (CMB) likelihoods [26]. Specifically, the low-ℓ likelihoods (`Commander`, `SimAll`) are computed at the pixel map level due to the non-Gaussian nature of the power spectrum at large scales; and the high-ℓ likelihoods (`Plik`) rely on "pseudo-C_ℓ's" from various frequency channels, which introduce 47 nuisance parameters to describe instrument noise and foregrounds. Such dataset is too complex to be simulated in large numbers, preventing its integration into SBI workflows and thus limiting the ability to leverage the benefits previously discussed.

In this work, we present a simple yet general trick to transform any explicit likelihood into an *effective simulator*. Our method introduces an *auxiliary-observable*, derived from samples of a pre-converged MCMC run, to construct an effective simulator that faithfully reproduces the original explicit likelihood. This effective simulator can be directly used for SBI, inheriting the full suite of benefits offered by such framework. Specifically, from a single MCMC, one can rapidly generate many effective simulations, which can be reused across different SBI tasks. Moreover, SBI allows seamless combination of likelihoods, whether explicit ones (turned into effective simulators) or implicit. This directly addresses the com-

[*] g.francoabellan@uva.nl

[1] To name just a few concrete examples, the recent DESI BAO DR2 analysis [7] explored 14 configurations of experiments to test the w_0w_aCDM model; the 'H_0 Olympics' paper [14] analyzed 12 data combinations across 8 different 'finalist' models; and 30 combinations of experiments were considered in [15] to derive forecast constraints on neutrino masses.

putational challenge of performing massive global scans, as it enables inference without the need for additional model evaluations each time new data is added.

Alternatively, our approach can be used to speed up traditional sampling-based inference, by leveraging our effective simulator to train a fast likelihood emulator. This connects naturally to prior work on emulation in cosmology, where surrogate models are trained to approximate computationally expensive theory codes [27–29] or likelihoods [30, 31] and then used as fast drop-in replacements in MCMC samplers (see also [32, 33] for different methods to accelerate joint cosmological analyses).

We validate our method through a series of cosmological applications. First, we build an effective *Planck* ΛCDM simulator and apply it within SBI, recovering posteriors that are in excellent agreement with the original MCMC. Next, we combine this effective simulator with a forecast simulator of 3×2pt probes (weak lensing, galaxy clustering, and their cross-correlation) for a Stage IV photometric galaxy survey, and perform a joint ΛCDM analysis. Thanks to simulation reuse enabled by SBI, we find accurate posteriors for this data combination without requiring any new model evaluations.

Finally, we demonstrate our approach on a directly relevant physics application, by extending the SBI analysis to probe dynamical dark energy parameterized by the Chevallier-Polarski-Linder (CPL) equation-of-state [34, 35]. This is motivated by the recent exciting results by the DESI collaboration, which hint at departures from a cosmological constant [36]. Specifically, the combination of data from CMB + DESI baryonic acoustic oscillation (BAO) + Type Ia supernovae (SNIa) shows a preference for CPL evolving dark energy over ΛCDM at the $\sim 3-4\sigma$ level, depending on the SNIa dataset used [7]. This has sparked a strong debate in the community about the robustness of the results and potential implications for beyond-ΛCDM models [37–50].

We explore the impact of Stage IV photometric surveys like *Euclid* in light of the latest DESI results, as these surveys are expected to achieve percent-level constraints on the dark energy equation-of-state [51]. To this end, we adopt the following strategy: we generate synthetic Stage IV photometric data assuming the CPL cosmology favored by present CMB+DESI BAO+SNIa data, and assess the resulting parameter uncertainties given different data configurations. Notably, we find that Stage IV photometric data alone could yield a $\sim 5\sigma$ detection of evolving dark energy, while its combination with current datasets could raise this significance to almost 7σ. This underscores the crucial role that joint analyses will play in advancing our understanding of dark energy.

This paper is organized as follows. In Sec. II, we present the methodology, describing the auxiliary-observable trick and the framework for combining explicit and implicit likelihoods. In Sec. III we detail the application to cosmology, starting with a description of the models and datasets considered, and then discussing the inference setup for both MCMC and SBI. In Sec. IV, we show our main results for the ΛCDM and CPL dark energy models. We conclude in Sec. V. Complementary information can be found in various appendices. In App. A we propose an alternative formulation of the auxiliary-observable trick. In App. B we describe an application of our trick to build likelihood emulators that can be used for fast joint cosmological analyses. In App. C we present the results of an empirical coverage test for the SBI results. Finally, in App. D we discuss our measure of the preference for evolving dark energy.

II. METHODOLOGY

A. The auxiliary-observable trick

Our goal is to embed a given likelihood function into a simulation-based framework. We begin with a likelihood function $L(\boldsymbol{\theta}) \equiv p(\boldsymbol{x}_o \mid \boldsymbol{\theta})$, defined for an observation $\boldsymbol{x}_o$, that we can evaluate for different parameter values $\boldsymbol{\theta}$. The objective is to construct an effective simulator for an auxiliary-observable, dubbed $\boldsymbol{a}$, that allows the use of SBI methods while preserving the information from our original likelihood and enabling inference on $\boldsymbol{x}_o$. The construction works as follows.

First, we define a probability distribution that is proportional to our likelihood function:

$$p_L(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta}). \tag{1}$$

This distribution essentially captures the shape of the likelihood in parameter space. In practice, samples from $p_L(\boldsymbol{\theta})$ can be obtained from previous Monte Carlo chains generated using the original likelihood $L(\boldsymbol{\theta})$ with wide uninformative priors over the parameters $\boldsymbol{\theta}$.

The crucial step is to define the probability distribution of the auxiliary-observable. We do so via

$$p(\boldsymbol{a} \mid \boldsymbol{\theta}) \equiv p_L(\boldsymbol{\theta} - \boldsymbol{a}) \ . \tag{2}$$

An important consequence of this definition is that, when evaluating the distribution at $\boldsymbol{a} = \boldsymbol{0}$, we recover our original likelihood up to a normalizing factor:

$$p(\boldsymbol{a} = \boldsymbol{0} \mid \boldsymbol{\theta}) \propto L(\boldsymbol{\theta}) \ . \tag{3}$$

This property is what enables our construction to work.

To generate samples from this auxiliary-observable distribution, we use a simple procedure. For any given parameter $\boldsymbol{\theta}$, we sample $\boldsymbol{a} \sim p(\boldsymbol{a} \mid \boldsymbol{\theta})$ by first drawing a random parameter $\boldsymbol{\theta}'$ from the distribution $p_L(\boldsymbol{\theta}')$ and then setting:

$$\boldsymbol{a} = \boldsymbol{\theta} - \boldsymbol{\theta}' \quad \text{with} \quad \boldsymbol{\theta}' \sim p_L(\boldsymbol{\theta}') \text{ and } \boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \ , \tag{4}$$

where $p(\boldsymbol{\theta})$ is our parameter prior. This sampling procedure yields pairs $(\boldsymbol{a}, \boldsymbol{\theta})$ that follow the joint distribution $p(\boldsymbol{a}, \boldsymbol{\theta}) = p(\boldsymbol{a} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$.

With these samples in hand, we can train any SBI algorithm – whether neural posterior estimation (NPE)

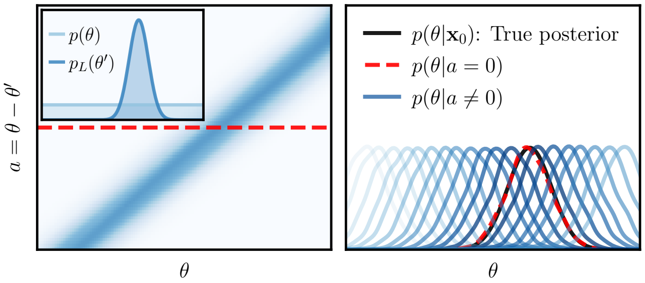

FIG. 1. Summary illustration of the auxiliary-observable construction for a toy 1-dimensional problem. Left panel: Definition of the auxiliary-observable a according to Eq. 4 and its relationship with θ. Right panel: Multiple conditional distributions $p(\theta \mid a)$ for different values of the auxiliary-observable a, with the true posterior $p(\theta \mid x_0)$ (black solid line) recovered for $a = 0$ (red dashed line). This shows how the framework preserves the original likelihood information at $a = 0$.

[52, 53], neural likelihood estimation (NLE) [54, 55], or neural ratio estimation (NRE) [56, 57]. Thanks to this construction, when the trained SBI estimators are evaluated at $\boldsymbol{a} = \mathbf{0}$, we recover the desired inference for the original observation $\boldsymbol{x}_o$. For instance, in the case of a NPE network $\hat{q}_\phi(\boldsymbol{\theta} \mid \boldsymbol{a})$, one obtains

$$\hat{q}_\phi(\boldsymbol{\theta} \mid \boldsymbol{a} = \mathbf{0}) \simeq p(\boldsymbol{\theta} \mid \boldsymbol{x}_o) \ . \tag{5}$$

Importantly, for a given dataset and cosmological model, a single MCMC run is sufficient to generate as many effective simulations as needed – whether for inference on the same dataset or in combination with others, as will be detailed in the next subsection.

A visual summary of the auxiliary-observable trick is presented in Fig. 1 for a simple 1-dimensional case, illustrating how this construction preserves the original likelihood information when evaluated at $\boldsymbol{a} = \mathbf{0}$. An alternative formulation of the auxiliary-observable trick that provides additional intuition is presented in App. A.

B. Combining explicit and implicit likelihoods

The auxiliary-observable construction opens up the possibility of combining explicit and implicit likelihoods within a unified SBI framework. Consider a scenario where we have both a simulation model $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ (implicit likelihood) with observation $\boldsymbol{y}_o$, and some additional likelihood constraint $L(\boldsymbol{\theta})$ (explicit likelihood) with observation $\boldsymbol{x}_0$ that we want to combine into our analysis. This explicit likelihood can be converted into an effective simulator for $\boldsymbol{a}$ following Eq. 4, thereby enabling the use of simulations for both data sources.

We can then sample from the combined model:

$$\boldsymbol{a}, \boldsymbol{y}, \boldsymbol{\theta} \sim p(\boldsymbol{a} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \ , \tag{6}$$

where both $\boldsymbol{a}$ and $\boldsymbol{y}$ serve as training samples in our SBI runs. At inference time, to obtain joint constraints for observations $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$, one just needs to evaluate the inference networks at $\boldsymbol{a} = \mathbf{0}$ and $\boldsymbol{y} = \boldsymbol{y}_0$.

It is worth noting that the combined simulation model in Eq. 6 assumes statistical independence between datasets, i.e. $p(\boldsymbol{a}, \boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{a} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta})$. This is usually a good approximation for many cosmological observables, and we adopt it throughout this work. As a real world application, in Sec. IV we will show examples combining the full *Planck* CMB likelihoods and a simulation model for a Stage IV photometric survey similar to *Euclid*.[2] We also note that, given the independence assumption, this combined analysis could also be performed by simply generating simulations of Stage IV data with *Planck* priors. However, we emphasize that our method additionally allows to have *Planck* and Stage IV as separate simulators, that can be reused for multiple data combinations (this will be illustrated in Fig. 3).

III. APPLICATION TO COSMOLOGY

A. Models and data

In this work, we analyze two different cosmological scenarios: the standard flat ΛCDM model, which assumes a cosmological constant dark energy with equation-of-state parameter $w = -1$; and the $w_0 w_a$CDM model, which assumes a time-evolving dark energy with the CPL [34, 35] parametrization for the equation-of-state parameter

$$w(a) = w_0 + w_a(1 - a). \tag{7}$$

To analyze the ΛCDM model we use the parameter basis $\{H_0, 100\omega_b, \omega_\mathrm{cdm}, n_s, \ln{(10^{10} A_\mathrm{s})}, \tau_\mathrm{reio}\}$, where H_0 is the Hubble constant, ω_b and ω_cdm are the physical baryon and cold dark matter abundances, A_s and n_s are the amplitude and tilt of the primordial power spectrum, and τ_reio is the reionization optical depth. For $w_0 w_a$CDM we additionally vary the parameters w_0 and w_a, and use the parametrized post-Friedmann (PPF) approach [60, 61] to handle phantom crossing. To compute the quantities needed for the different cosmological observables (e.g., Hubble parameter $H(z)$, matter power spectrum $P_m(k, z)$, CMB anisotropy spectra C_ℓ^{XY}) we use the public Boltzmann solver `CLASS`[3] [13, 62] with the **`halofit`** prescription [63, 64] for the non-linear corrections to the matter power spectrum. We adopt the *Planck* convention in modeling free-streaming neutrinos as two massless species and one massive with $m_\nu = 0.06$ eV [1].

[2] In principle, there exist correlations between CMB and large-scale structure (LSS) observables, arising mostly from secondary anisotropies of the CMB photons. We neglect these correlations as done in previous CMB+*Euclid* forecasts (e.g. [58]), but we note that this correlation will be more important for a survey very sensitive to CMB lensing like CMB-S4 [59].

[3] https://github.com/lesgourg/class_public

z_m	$\bar{n}_g$ [arcmin^{-2}]	$f_{\rm sky}$	$\Delta \log_{10} \ell$	σ_ϵ	$f_{\rm out}$
0.9	30	0.35	0.08	0.3	0.1
c_b	z_b	σ_b	c_o	z_o	σ_o
1.0	0	0.05	1.0	0.1	0.05

TABLE I. Parameters describing the specifications for our Stage IV photometric survey. These include the median redshift of the survey z_m, the surface galaxy density $\bar{n}_g$, the sky fraction $f_{\rm sky}$ and the intrinsic ellipticity error σ_ϵ. For definitions of the remaining parameters, we refer the reader to [24].

We will confront the ΛCDM and w_0w_aCDM models against real data from CMB, BAO and SNIa, as well as synthetic data from an upcoming Stage IV photometric survey. We describe each of these in the following subsections.

1. CMB, BAO and SNIa datasets

For CMB, BAO and SNIa, we use the datasets detailed in the bullet points below.

- ***Planck* 2018**: We consider the temperature (TT), polarization (EE) and cross (TE) power spectra from *Planck*, specifically using the `Commander`, `SimAll` (for multipoles $\ell < 30$) and `Plik` (for multipoles $\ell \geq 30$) likelihoods, together with the lensing amplitude reconstruction from the official PR3 data release [26, 65]. These likelihoods introduce 47 nuisance parameters to model instrument noise and foregrounds, of which we vary 21 following `Plik` recommendations.

- **DESI BAO DR2**: We make use of the DESI BAO DR2 distance measurements from Table IV in [7]: D_V/r_d at $z = 0.295$; D_M/r_d and D_H/r_d (including their correlation) at $z = 0.51$, 0.706, 0.934, 1.321, 1.484 and 2.33.

- **PantheonPlus**: We use the Pantheon+ catalog, which compiles information about the luminosity distance to over 1550 SNIa in the redshift range $0.001 < z < 2.3$ [66, 67]. This likelihood introduces one nuisance parameter $\mathcal{M}$ which describes the SNIa calibration.

In the following, we will use the data combination '*Planck* 2018 + DESI BAO DR2 + PantheonPlus' (hereinafter referred to as 'Baseline') to put constraints on w_0w_aCDM model, whereas for ΛCDM model we will omit the DESI BAO DR2 and PantheonPlus datasets. This choice is motivated by the mild $\sim 2.3\sigma$ tension reported by DESI [7] between their BAO data and the CMB within the ΛCDM model, a discrepancy which is alleviated for the w_0w_aCDM model.

2. Synthetic 3×2pt data

On top of our *Planck*/Baseline datasets, we consider data from a simulated **Stage IV photometric survey**. In particular, we focus on the so-called 3×2pt signal, i.e. the combination of weak lensing (WL), photometric galaxy clustering (GCph), and their cross-correlation. This is described by a series of angular power spectra $C_{ij}^{AB}(\ell)$, where i and j label different tomographic redshift bins, and A and B refer to either WL or GCph. For the construction of such observables from `CLASS` outputs, we use the same modeling approach as in [24], and we refer the reader to that work for a detailed description.

We consider $N_z = 10$ tomographic bins, and compute each spectrum $C_{ij}^{AB}(\ell)$ for $N_\ell = 29$ log-spaced values between $\ell_{\rm min} = 10$ and $\ell_{\rm max} = 2000$.[4] We assume these spectra to be distributed according to a multivariate Gaussian distribution; the parameters used to compute the data covariance matrix and the photometric redshift distributions are given in Tab. I (these correspond to the specifications for a Stage IV survey like *Euclid*). Our modeling of the 3×2pt spectra necessitates the variation of 12 nuisance parameters: $A_{\rm IA}, \eta_{\rm IA}$, which describe intrinsic alignments, and $b_1, ..., b_{10}$, which describe photometric galaxy bias (one per redshift bin).

To generate our synthetic dataset, we also need a fiducial cosmological model that we assume to be the one describing the Universe. Since we aim to combine mock 3×2pt data with actual measurements from CMB, BAO and SNIa, this choice must be done with care – particularly because the maximum of the combined likelihoods may lie far from the fiducial point. Hence, we follow a similar strategy as the 'fitted-Fisher approach' of [69], and generate two sets of mock 3×2pt data: one from the best-fit of *Planck* (used for the ΛCDM analysis) and another from the best-fit of our Baseline data (used for the w_0w_aCDM analysis). The corresponding fiducial values are reported in Tab. II. To obtain the best-fit parameters, we minimize the χ^2 with the method explained in Appendix D.1 of [14]. For visualization purposes, we assume a 'noiseless' 3×2pt observation, so that we obtain posteriors that are centered on the fiducial values.

B. Inference setup

We will confront each model against three different data combinations:

- ΛCDM: [*Planck*, 3×2pt, *Planck*+3×2pt],
- w_0w_aCDM: [Baseline, 3×2pt, Baseline+3×2pt].

[4] We note that our choice $\ell_{\rm max}^{\rm GC_{ph}} = \ell_{\rm max}^{\rm WL} = 2000$ lies between the 'pessimistic' ($\ell_{\rm max}^{\rm GC_{ph}} = 750$, $\ell_{\rm max}^{\rm WL} = 1500$) and 'optimistic' ($\ell_{\rm max}^{\rm GC_{ph}} = 3000$, $\ell_{\rm max}^{\rm WL} = 5000$) configurations that have been usually considered in previous *Euclid* forecasts (e.g. [58, 68, 69]).

Parameter	Fiducial I	Fiducial II
H_0 [km s^{-1} Mpc^{-1}]	67.417	67.904
$100\,\omega_{\rm b}$	2.2392	2.2482
$\omega_{\rm cdm}$	0.1199	0.1190
$n_{\rm s}$	0.9667	0.9691
$\ln(10^{10}A_{\rm s})$	3.0452	3.0448
w_0	-1.0	-0.8276
w_a	0.0	-0.6653
$A_{\rm IA}$	1.72	1.72
$\eta_{\rm IA}$	-0.41	-0.41
b_1	1.0998	1.0998
b_2	1.2203	1.2203
b_3	1.2724	1.2724
b_4	1.3166	1.3166
b_5	1.3581	1.3581
b_6	1.3998	1.3998
b_7	1.4447	1.4447
b_8	1.4965	1.4965
b_9	1.5653	1.5653
b_{10}	1.7430	1.7430

TABLE II. Fiducial cosmological and nuisance parameters values used to generate the mock 3×2pt data. The cosmological parameters of 'Fiducial I' correspond to the ΛCDM best-fit of *Planck* 2018 data, while those of 'Fiducial II' correspond to the w_0w_aCDM best-fit of our Baseline dataset (*Planck* 2018 + DESI BAO DR2 + PantheonPlus). The values for the nuisance parameters are the same as in [24].

We perform each of these six different analyses using both MCMC and SBI (with the help of the auxiliary-observable trick). For the MCMC part, we use the public code `MontePython-v3`[5] [70, 71], where all the likelihoods described in the previous subsection have been implemented. We employ a Metropolis-Hastings (MH) algorithm assuming wide flat priors[6] on the various cosmological parameters. We deem chains to be converged with the Gelman-Rubin [72] criterion $|R-1| \leq 0.03$, and produce figures via the `GetDist`[7] package [73]. For the SBI analysis presented in the main body of the paper, we use an algorithm called Marginal Neural Ratio Estimation (MNRE) [74], implemented via the open-source code `Swyft`[8]. MNRE uses simulated data-parameter pairs to train neural classifiers that directly learn 1-dimensional and 2-dimensional marginal posteriors of interest (see e.g. [75–81] for astrophysical or cosmological use cases). Like other SBI methods, MNRE requires two key components for each dataset: a forward simulator and a neural network design. The network is typically divided into two parts: the first performs a compression of the data into a small number of features, and the second carries out the actual inference based on the feature-parameter pairs. In the following, we describe these components for both the 3×2pt and *Planck*/Baseline datasets.

1. SBI pipeline for 3×2pt data

To perform inference from a Stage IV photometric survey, we employ the same forecast simulator of 3×2pt power spectra that was used by [24]. We note that this simulator samples data vectors from a multivariate Gaussian with given covariance, making it equivalent to the explicit Gaussian likelihood that we incorporated in `MontePython-v3`. We do this for simplicity and to be able to cross-check our SBI results against MCMC, but we stress that the framework presented here would also allow to consider a more advanced simulator capturing any non-Gaussianities in the likelihood, like the one developed in [19].

We generate 5×10^4 simulations of 3×2pt photometric spectra by varying the cosmological parameters $\{H_0, 100\omega_b, \omega_{\rm cdm}, n_s, \ln(10^{10}A_{\rm s})\}$[9] as well as the nuisance parameters $\{A_{\rm IA}, \eta_{\rm IA}, b_1, ..., b_{10}\}$. When testing CPL dark energy, we generate another set of 5×10^4 simulations where we additionally vary $\{w_0, w_a\}$. Each of these simulation batches was generated in parallel in less than 2 hours using 72 CPU cores. We define the prior region based on the results of two preliminary Fisher analyses of 3×2pt (one for ΛCDM and another for w_0w_aCDM), following a similar approach to that of [24]. In particular, each parameter θ_i is drawn from a uniform prior

$$\theta_i \sim \mathcal{U}([\theta_i^0 - 3\sigma_i^F, \theta_i^0 + 3\sigma_i^F]), \tag{8}$$

where θ_i^0 denotes the fiducial values in Tab. II and the errors σ_i^F are estimated from the corresponding Fisher matrix F_{ij} as $\sigma_i^F = \sqrt{(F^{-1})_{ii}}$. For the parameters w_0 and w_a, we adopt slightly broader priors,

$$\theta_i \sim \mathcal{U}([\theta_i^0 - 5\sigma_i^F, \theta_i^0 + 5\sigma_i^F]). \tag{9}$$

The priors in Eq. 8 might look relatively narrow from the point of view of a 3×2pt-only analysis, but these are usually much broader than *Planck*-based posteriors (as we will see in Fig. 2). We also remark that very wide priors can easily be accommodated in SBI using sequential methods, which apply active learning techniques to quickly zoom into the relevant region of the parameter space [54, 75].

[5] https://github.com/brinckmann/montepython_public

[6] We did not find the choice of prior boundaries to be particularly relevant for MCMC (as long as the posteriors never hit the prior), since the MH algorithm always explores points in the vicinity of the best-fit.

[7] https://github.com/cmbant/getdist

[8] https://github.com/undark-lab/swyft

[9] It should be noted that *Euclid* probes are insensitive to $\tau_{\rm reio}$.

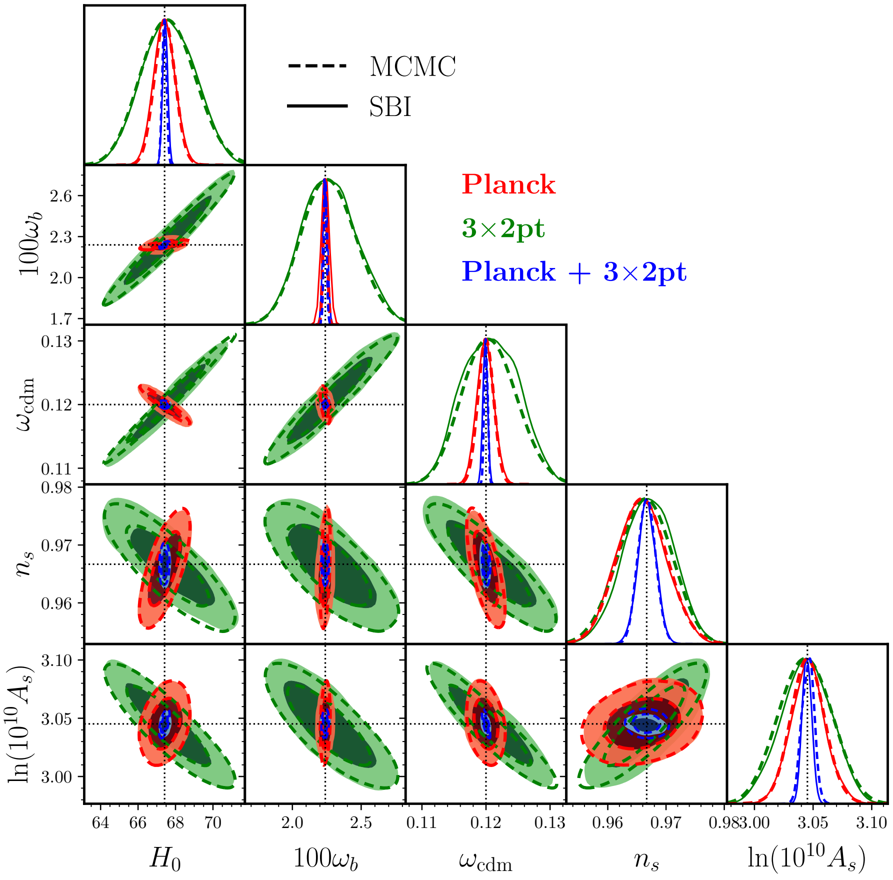

FIG. 2. 1- and 2-dimensional marginalized posterior distributions (68% and 95% C.L.) of the ΛCDM cosmological parameters for different data combinations, using both MCMC (dashed lines) and SBI (solid lines). The black dotted lines indicate the ΛCDM best-fit to *Planck* 2018, which is our 'Fiducial I' model used to generate the mock 3×2pt observation for a Stage IV photometric survey. SBI posteriors are in excellent agreement with MCMC.

For the data reduction of 3×2pt power spectra, we employ the same strategy as in [24], which combines Cholesky decomposition, principal component analysis (PCA), and linear compression to produce parameter-specific data summaries. These features are concatenated with the corresponding parameters, and fed as input to the ratio estimators, in charge of estimating every 1- and 2-dimensional marginal posterior. Throughout this work, each ratio estimator is implemented as a multi-layer perceptron (MLP) consisting of four residual blocks with 128 neurons each.

2. *SBI pipeline for Planck/Baseline datasets*

For the *Planck* and Baseline datasets, we construct effective simulators using the trick described in Sec. II A. Namely, we generate samples of the auxiliary-observable $\boldsymbol{a}$ simply by drawing parameters $\boldsymbol{\theta}$ from the prior and substracting random samples $\boldsymbol{\theta}'$ from the corresponding MCMC run. To get each sample $\boldsymbol{\theta}'$, we read the accepted points and weights stored in the MCMC files (excluding the burn-in phase), and then randomly select a point using the weights as probabilities. In this way, we build an effective ΛCDM simulator for *Planck*, and an effective w_0w_aCDM simulator for the Baseline combination.[10]

We generate 10^5 samples from our effective *Planck* and Baseline simulators, by varying the cosmological parameters with the same priors as in Eq. 8-Eq. 9. We emphasize that these samples can be generated almost instantaneously, since they just rely on samples from a pre-converged MCMC.

The auxiliary samples for *Planck*/Baseline have already the same dimensionality as the parameters of interest, so the pre-compression step is not required. However, we still found it useful to standardize the samples using

[10] The auxiliary-observable trick is technically not necessary for the DESI BAO DR2 and PantheonPlus datasets, as these are described by simple Gaussian likelihoods, which a priori could be easily reformulated as simulators. However, since we decided to analyze these likelihoods always in combination with *Planck*, we found easier to apply the trick to the full Baseline dataset.

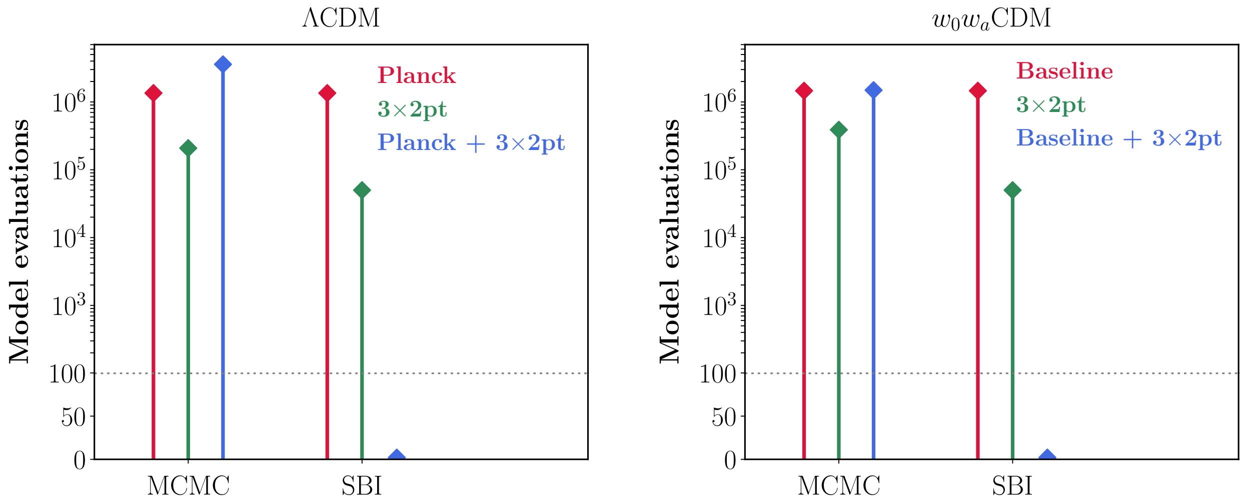

FIG. 3. Number of required model evaluations across all cosmological scenarios, datasets and inference methods considered in this work. For MCMC, this reflects the number of likelihood evaluations needed for convergence, whereas for SBI, the meaning is different for each simulator. On the one hand, our effective simulators of *Planck* 2018 and Baseline (= *Planck* 2018 + DESI BAO DR2 + PantheonPlus) datasets use 10^5 posterior samples that can be promptly generated from the corresponding pre-converged MCMC runs, and hence necessitate the same number of evaluations as MCMC. On the other hand, our forward 3×2pt simulators use only 5×10^4 data realizations, a factor ~ 6 smaller than those needed for MCMC. Finally, for the SBI combined analyses, we reused the samples from the *Planck*/Baseline and 3×2pt runs, hence requiring *zero* new model evaluations. All SBI runs additionally involve training the inference networks, which takes just $5-20$min on a single GPU.

online normalization [82]. When incorporating the effective *Planck*-based simulators in the combined analyses, we simply concatenate the auxiliary samples with the features extracted from the 3×2pt spectra within each marginal.

3. *Training strategy*

For all the SBI runs carried out in this work, we allocate 80% of the simulations for training, and the remaining 20% for validation. We use the Adam optimizer with a batch size of 256 and an initial learning rate of 10^{-3}. The learning rate is reduced by a factor of 0.1 if the validation loss plateaus for 3 consecutive epochs, and training is stopped if no improvement is observed for 5 epochs. Training for each inference run takes $\sim 5-20$ min on a single 40 GB `Nvidia A100` GPU.

IV. RESULTS

A. ΛCDM analysis

In Fig. 2 we show the cosmological constraints from the ΛCDM analysis of *Planck* 2018, 3×2pt Stage IV photometric data, and their combination. We compare the results obtained with MCMC (dashed) against those obtained with SBI (solid). The dotted lines indicate the ΛCDM best-fit to *Planck*, used to generate the mock 3×2pt observation. In Fig. 3 we additionally show the number of model evaluations needed for all the analyses carried out in this work.

We find that the SBI and MCMC posteriors are always in excellent agreement, with both methods yielding unbiased parameter constraints. Importantly, the agreement observed at the level of *Planck* contours (red) confirms the effectiveness of the auxiliary-observable trick that we introduced in Sec. II, that is, our effective *Planck* simulator preserves the original likelihood information.

A natural caveat is that constructing our effective simulator requires having first the corresponding MCMC runs. However, it is crucial to emphasize that this step needs to be performed only once per theoretical model, after which the resulting simulator can be reused for multiple data combinations. In practice, this permits a massive reduction in the number of simulations for combined analyses. Namely, for our SBI combined analysis (blue dashed), we could recycle both the 5×10^4 simulations of 3×2pt spectra (already a factor ~ 6 smaller than those needed for MCMC) as well as the *Planck* effective samples, hence allowing us to perform inference with *zero* additional model evaluations. On the contrary, the MCMC combined analysis (blue solid) necessitated roughly $\sim 3 \times 10^6$ likelihood evaluations to achieve convergence.[11]

[11] Among all the MCMC runs we performed, the combined *Planck*+3×2pt ΛCDM analysis was notably slower to converge due to a poor prior knowledge of the covariance matrix used for the initial proposal distribution.

Data set	MCMC	SBI
Baseline	3.31σ	3.29σ
3×2pt	5.0σ	4.99σ
Baseline+3×2pt	6.80σ	6.82σ

TABLE III. Significance of the detection of w_0w_aCDM relative to ΛCDM, estimated as the Mahalanobis distance $Q_{\boldsymbol{\theta}}$ between our fiducial point $(w_0, w_a) = (-0.8276, -0.6653)$ and $(w_0, w_a) = (-1, 0)$, for the various data sets and methods considered in this work.

This demonstrates how our trick enables the combination of explicit and implicit likelihoods in a fast and simulator-efficient manner, without ever needing a full generative model of the likelihood. Moreover, doing inference within SBI brings additional advantages, like the possibility to perform statistical consistency checks which are usually unfeasible for MCMC. We illustrate this point in App. C, where we conduct an empirical coverage test to further check the statistical consistency of the trained network.

Even if our trick was originally designed to integrate arbitrary likelihoods into SBI, we demonstrate in App. B that it can alternatively be used to construct likelihood emulators for accelerating joint MCMC analyses.

B. w_0w_aCDM analysis

We now turn to the case of CPL evolving dark energy. As we mentioned in Sec. I, our goal is to find the significance with which Stage IV photometric surveys such as *Euclid* could detect the w_0w_aCDM cosmology favored by the combination of DESI BAO, CMB and SNIa data. We also want to show how such kind of joint analysis, traditionally carried out with sampling-based methods, can be done much more efficiently with SBI, thanks to our auxiliary-observable trick.

In Fig. 4, we show the cosmological constraints on the w_0-w_a plane for the different data combinations and the different methods (MCMC, SBI) considered in this paper. The star indicates the w_0w_aCDM best-fit to our Baseline data, used to create the synthetic 3×2pt observation. We observe that SBI posteriors are again in excellent agreement with MCMC. As for ΛCDM, we could exploit simulation reuse to perform the Baseline+3×2pt combined analysis without needing any additional model evaluations (this is illustrated in the right panel of Fig. 3).

Regarding the scientific conclusion, our Baseline dataset indicates a preference for the w_0w_aCDM model over ΛCDM at the $\sim 3.3\sigma$ level (see App. D for details about our measure of the statistical significance of preference for evolving dark energy). This is comparable to the $\sim 2.8\sigma$ preference reported by the DESI collaboration using similar datasets[12] [7]. Interestingly, we find that 3×2pt data from an *Euclid*-like survey alone could detect the best-fit w_0w_aCDM model at the $\sim 5\sigma$ level, while the full combination of all data could raise the detection to the $\sim 6.8\sigma$ level. We remark that these numbers may slightly vary depending on the choice of $\ell_{\rm max}$ or the non-linear prescription applied to model weak lensing cosmic shear [68]. However, we expect that our main conclusion, namely that Stage IV photometric measurements can detect dynamical dark energy hinted by DESI with high significance, to remain valid under small variations in the 3×2pt modeling. We also emphasize that it is the full combination of cosmic shear and photometric galaxy clustering spectra that provides sufficient constraining power to robustly measure dynamical dark energy. Indeed, it was recently shown in [85] that Stage IV cosmic shear alone could distinguish evolving dark energy preferred by DESI only at the field-level, but not at the power spectrum level.

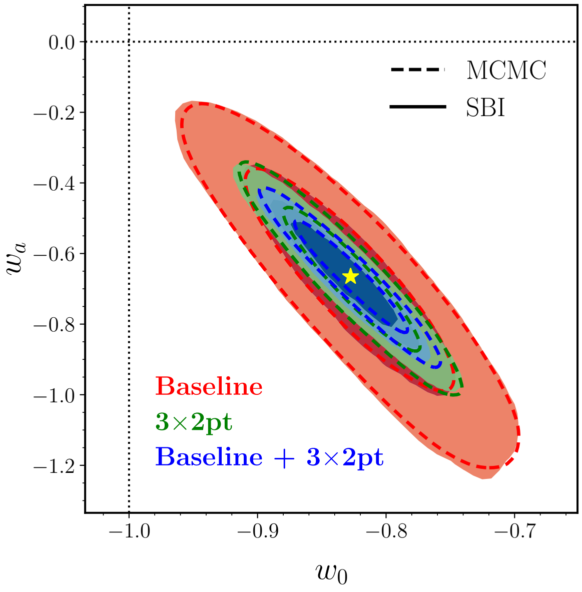

FIG. 4. 2-dimensional marginalized posterior (68% and 95% C.L.) of w_a and w_0 for different data combinations, using both MCMC (dashed lines) and SBI (solid lines). The star indicates the w_0w_aCDM best-fit to our Baseline dataset (*Planck* 2018 + DESI BAO DR2 + PantheonPlus), which is our 'Fiducial II' model used to generate the mock Stage IV 3×2pt observation. The black dotted lines indicate $w_0 = -1$ and $w_a = 0$; the ΛCDM limit lies at their intersection. The significance of rejection of ΛCDM is 3.3σ, 5.0σ and 6.8σ for Baseline, 3×2pt and their combination, respectively.

[12] We note that the baseline analysis by the DESI collaboration relies on newer CMB data. In particular, instead of using `Plik`, they adopt `Camspec` [83], which is built on the `NPIPE` PR4 data release from *Planck*, along with CMB lensing data from ACT DR6 [84]. However, the constraints on the w_0-w_a plane were found very consistent using `Camspec` and `Plik` [7].

V. CONCLUSIONS AND OUTLOOK

In this paper, we have introduced a simple yet effective way to incorporate arbitrary explicit likelihoods into modern simulation-based analysis frameworks. This is particularly relevant for legacy cosmological likelihoods such as *Planck*, which cannot be easily reformulated as physical forward simulators. The approach, presented in Sec. II, is based on constructing effective simulators from pre-converged MCMC runs, enabling the combination of explicit and implicit likelihoods within a unified SBI analysis pipeline. Alternatively, these effective simulators can be used to accelerate joint constraints in classical sampling-based inference by training fast likelihood emulators, as shown in App. B.

Importantly, for a given theoretical model, a single converged MCMC is needed to define such an effective simulator, which can subsequently be reused across different data combinations. Moreover, converged chains are routinely released by major survey collaborations alongside their data products (see e.g. [86]), providing a ready resource for constructing effective simulators without requiring access to the original likelihood code. This makes it possible to perform global fits very efficiently, bypassing expensive model evaluations for every new experimental configuration.

We validated our method by conducting several cosmological applications, in which each analysis was performed using both MCMC and SBI, and the resulting posteriors were compared. First, we did the analysis for the standard ΛCDM model, using the full *Planck* 2018 likelihoods, synthetic 3×2pt measurements for a Stage IV photometric survey, and the combination of both datasets. Within SBI, the *Planck* and 3×2pt probes were modeled as effective and forward simulators, respectively. We found excellent agreement between MCMC and SBI (shown in Fig. 2), showcasing the effectiveness of our method for combining forward simulators with explicit likelihoods, even when a generative model of the latter is not available. Furthermore, we could exploit simulation reusability to perform the SBI joint analysis at *zero* simulator cost, greatly reducing computational time.

Next, we applied the previous analysis setup to the w_0w_aCDM evolving dark energy model, and extended *Planck* with BAO data from DESI DR2 and SNIa data from PantheonPlus (constituting our baseline dataset). Again, we found very good overlap between SBI and MCMC results (shown in Fig. 4), thus demonstrating the applicability of our method to test ΛCDM extensions in a much more simulation-efficient way than MCMC. Our results indicate that Stage IV 3×2pt data alone could potentially raise the detection of evolving dark energy hinted by DESI from $\sim 3\sigma$ to $\sim 5\sigma$, and up to $\sim 7\sigma$ when combined with the baseline dataset.

Several clear improvements could extend the reach of our proposed framework. For instance, the current approach assumes statistical independence between datasets, which is a relatively common approximation in cosmological global fits. However, correlations are known to exist between certain cosmological probes, which may affect joint constraints. For future experiments, a way forward would be that major collaborations release realistic joint simulators for their various observables, such that correlations are automatically modeled. For past observations, one could simply construct joint likelihoods that properly account for the cross-correlations, and then apply the auxiliary-observable trick to preserve this structure within SBI analyses.

Another caveat is that our effective simulators require separate Monte Carlo chains for different theoretical models. This could quickly become a burden if one wants to explore a large number of cosmological scenarios, which has become a growing practice given the accumulation and persistence of observational discrepancies in cosmology (see e.g. [87] for a recent review). Hence, developing more model agnostic techniques could help reduce this computational overhead.

Nonetheless, the framework represents a significant advance, as it brings legacy data into alignment with the capabilities of modern SBI techniques, and it allows efficient combinations of multiple cosmological datasets.

ACKNOWLEDGMENTS

We thank Fabian Schmidt for his helpful comments on the draft of the manuscript. GFA, OS and CW acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 864035 - Undark).

Appendix A: Alternative formulation

We present here an alternative formulation of the auxiliary-observable trick that provides additional insight into the construction. Rather than working directly with the auxiliary-observable $\boldsymbol{a}$ as in the main text, this formulation frames the problem in terms of a simple simulator model with additive noise.

Our goal remains to embed a given likelihood function $L(\boldsymbol{\theta})$ into a simulation-based framework. To this end, we define an auxiliary simulator model $p(\mathbf{y} \mid \boldsymbol{\theta})$, and observed data $\mathbf{y}_{\text{obs}}$, such that $p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) \propto L(\boldsymbol{\theta})$. For a given likelihood function, many different simulator models exist that fulfill this requirement.

We focus here on the arguably simplest version of an auxiliary simulator model, which features homoscedastic (parameter-independent) noise $\mathbf{n} \sim p_N(\mathbf{n})$, and a trivial dependence of $\mathbf{y}$ to the model parameter $\boldsymbol{\theta}$. That model takes the form

$$\mathbf{y} = \boldsymbol{\theta} + \mathbf{n} \,. \tag{A1}$$

In order to relate $\mathbf{y}_{\text{obs}}$ and $p_N(\mathbf{n})$ to the likelihood function, we define a probability distribution that is proportional to our likelihood function:

$$p_L(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta}) \ . \tag{A2}$$

If we then define the auxiliary-observable as the mean value of samples from $p_L(\boldsymbol{\theta})$,

$$\mathbf{y}_{\rm obs} \equiv \mathbb{E}_{p_L(\boldsymbol{\theta})}[\boldsymbol{\theta}] \tag{A3}$$

and the (homscedastic) noise as centered and inverted samples from p_L,

$$p_N(\mathbf{n}) \equiv p_L(\mathbf{y}_{\rm obs} - \mathbf{n}) \ , \tag{A4}$$

it is straightforward to show, by using the definitions of the various distributions, that

$$p(\mathbf{y} = \mathbf{y}_{\rm obs} \mid \boldsymbol{\theta}) \propto L(\boldsymbol{\theta}) \ . \tag{A5}$$

This is the desired result.

Given samples from $p_N(\mathbf{n})$, we can generate samples from $p(\mathbf{y} \mid \boldsymbol{\theta})$ for arbitrary values of $\boldsymbol{\theta}$. Those noise samples, as well as the value of $\mathbf{y}_{\rm obs}$, can be generated using MCMC techniques. We first generate MCMC samples for $\boldsymbol{\theta}$ from $p_L(\boldsymbol{\theta})$. Based on those, we (a) estimate $\mathbf{y}_{\rm obs}$ (the mean value of $\boldsymbol{\theta}$) and (b) generate noise samples $\mathbf{n} \sim p_N(\mathbf{n})$ by centering the MCMC samples around zero (subtracting $y_{\rm obs}$).

With these samples in hand, we can train any SBI algorithm, as described in the main text. Thanks to this construction, when we evaluate the trained SBI estimators at $\mathbf{y} = \mathbf{y}_{\rm obs}$, we recover the desired inference for our original observation $\boldsymbol{x}_o$. For instance, in the case of a NPE network $\hat{q}_\phi(\boldsymbol{\theta} \mid \mathbf{y})$, one obtains

$$\hat{q}_\phi(\boldsymbol{\theta} \mid \mathbf{y} = \mathbf{y}_{\rm obs}) \simeq p(\boldsymbol{\theta} \mid \boldsymbol{x} = \boldsymbol{x}_o) \ . \tag{A6}$$

Appendix B: Using the effective simulations to build a likelihood emulator

As briefly mentioned in Sec. I, a natural application of our effective simulator is to train a fast likelihood emulator for sampling-based techniques. This approach offers a complementary pathway to the full SBI framework described in the main text, allowing us to leverage the computational benefits of the auxiliary-observable method – such as avoiding costly Boltzmann solver calls – while working with traditional MCMC inference pipelines. In this appendix, we demonstrate this application through a combined ΛCDM analysis of *Planck* and Stage IV 3×2pt data, in a setup identical to the one described in Sec. III.

Our approach involves training separate normalizing flows to emulate the likelihood for each dataset using NLE. For the *Planck* likelihood $p(\boldsymbol{x}_0 \mid \boldsymbol{\theta})$, we train a normalizing flow $\hat{q}_\phi(\boldsymbol{a} \mid \boldsymbol{\theta})$ on effective simulations $(\boldsymbol{a}, \boldsymbol{\theta})$ constructed from a previous *Planck* MCMC run as described in Sec. II A. Similarly, for the Stage IV 3×2pt likelihood $p(\boldsymbol{y}_0 \mid \boldsymbol{\theta})$, we train a second normalizing flow $\hat{q}_\phi(\boldsymbol{b} \mid \boldsymbol{\theta})$ on effective simulations $(\boldsymbol{b}, \boldsymbol{\theta})$ constructed from a former 3×2pt MCMC run. The mock 3×2pt observation $\boldsymbol{y}_0$ is the same that we used in Sec. IV A.

We use the package `sbi`[13] [88] to train the normalizing flows. For both flows, we employ a Masked Autoregressive Flow [89] architecture with 64 hidden features and 3 transformations. We use a learning rate of 10^{-3} and a batch size of 256, training on 10^5 effective simulations for each dataset. To standardize the input we use online normalization [82]; without this normalization, the normalizing flows struggles to learn the likelihood surfaces accurately enough.

Once trained, the likelihood emulators are obtained by evaluating the networks at the zero value of the auxiliary-observable. This evaluation recovers the original likelihood functions: $\hat{q}_\phi(\boldsymbol{a} = \mathbf{0} \mid \boldsymbol{\theta}) \simeq p(\boldsymbol{x}_o \mid \boldsymbol{\theta})$ for *Planck* and $\hat{q}_\phi(\boldsymbol{b} = \mathbf{0} \mid \boldsymbol{\theta}) \simeq p(\boldsymbol{y}_o \mid \boldsymbol{\theta})$ for the mock 3×2pt data. The combined log-likelihood is then constructed as the sum of the two individual log-likelihood emulators, assuming independence between the CMB and LSS datasets as discussed in the main text (see Sec. II B).

For sampling this joint likelihood, we use the `Nautilus` nested sampler [90]. Our results, shown in Fig. 5, indicate that the posterior contours using the emulated likelihoods (solid purple lines) overlap almost perfectly with the original MCMC (dashed blue lines), while taking only a fraction of the total CPU time. This excellent agreement validates both the accuracy of our auxiliary-observable construction for different SBI algorithms, as well as the reliability of the trained emulators.

Our likelihood emulation approach differs from most existing methods in cosmology, which typically focus on accelerating theoretical predictions like the matter power spectrum or CMB power spectra that feed into likelihood calculations (e.g., [27–29]). In contrast, our method directly emulates 'nuisance-marginalized' likelihoods from effective simulations, similar to the method proposed in [31]. Related techniques to accelerate joint likelihood-based analyses can be found in [32, 33].

Appendix C: Coverage tests

A key advantage of many SBI methods is that the trained inference networks are *amortized*, i.e. they can estimate the posterior not only for the actual observation (as in MCMC), but also for *any* mock observation simulated from the prior. This allows to asses fundamental statistical properties such as the *empirical coverage* [91], which gives the proportion of time that a given interval contains the true parameter value. For a well-calibrated posterior, the $p\%$ credible interval should contain the simulation-truth value $p\%$ of the time, so one should get a perfectly diagonal line when plotting the empirical

[13] https://github.com/sbi-dev/sbi

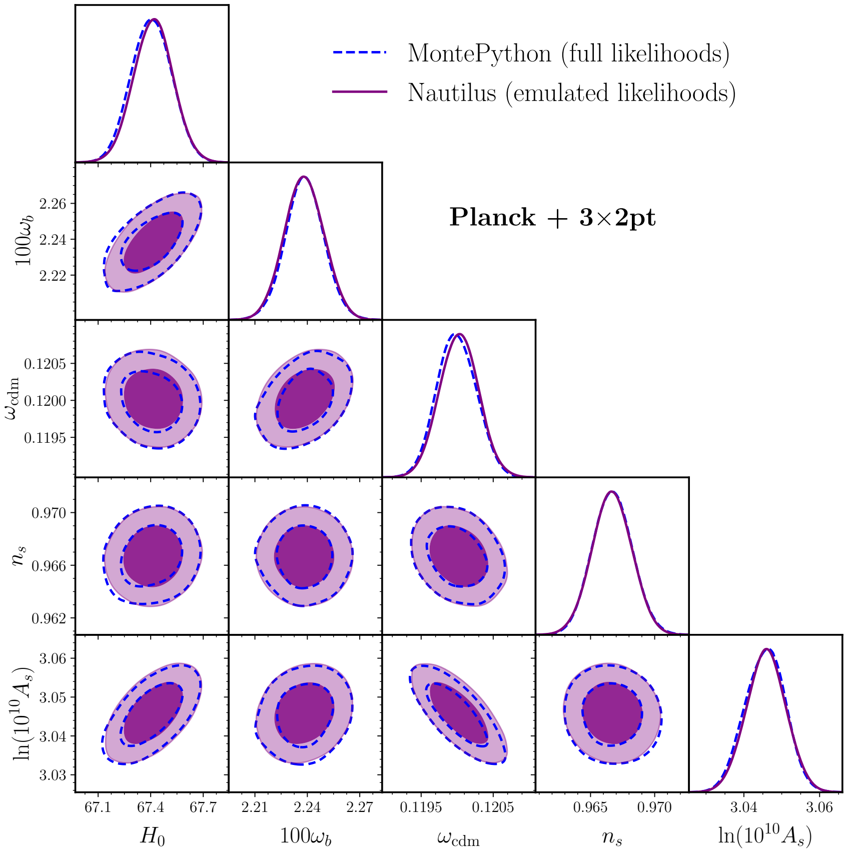

FIG. 5. 1- and 2-dimensional marginalized posterior distributions (68% and 95% C.L.) of the ΛCDM cosmological parameters for the combination of *Planck* and mock 3×2pt data for a Stage IV photometric survey. These were obtained using both the MCMC sampler `MontePython-v3` with the full likelihoods (dashed blue lines) and the nested sampler `Nautilus` with the emulated likelihoods (solid purple lines). The posteriors from the emulated likelihoods are in excellent agreement with MCMC.

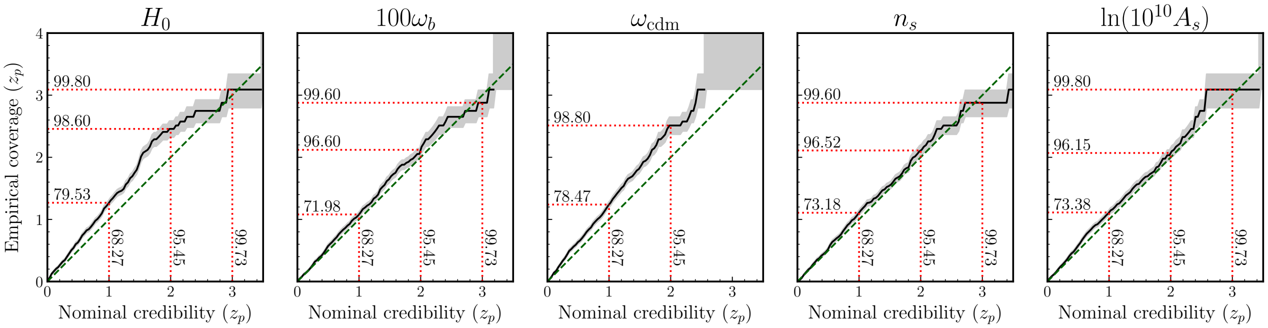

FIG. 6. Coverage test for the cosmological parameters from the ΛCDM analysis of *Planck*+3x2pt Stage IV photometric probes. The black solid line shows the average coverage, while the gray shaded region corresponds to the 68% uncertainty on the coverage. The empirical coverage and confidence level generally match to good precision.

coverage against the expected credible levels. Crucially, our auxiliary-observable framework enables such coverage tests because it provides an effective simulator that allows training of amortized posteriors. Once trained on effective simulations, these posteriors can be evaluated at any mock auxiliary-observable. Such tests are generally unfeasible for standard likelihood-based methods, where one would need to run hundreds of separate MCMC runs – one per each mock data simulation used in the coverage plot.

We perform a coverage test for the marginal ΛCDM posteriors of the *Planck*+3×2pt analysis, using a batch of 500 simulations. Instead of showing the highest posterior density region p, we use a different variable

z_p defined by $p/100 = \frac{1}{\sqrt{2\pi}}\int_{-z_p}^{z_p} dz \exp(-z^2/2)$, to put more emphasis in the tail of the posteriors. Hence, the usual $(1,2,3)\sigma$ regions correspond to $z_p = (1,2,3)$ with $p = (68.27, 95.45, 99.97)$. In addition, we compute the uncertainty on the empirical coverage arising from the finite number of samples, using the Jeffreys interval [75]. The results of the coverage test are shown in Fig. 6. We find that the empirical coverage and the confidence levels generally match to good precision.

Appendix D: Statistical significance of preference for evolving dark energy

We aim to quantify the preference of CPL dark energy across the various data combinations considered in this work. Model selection criteria such as the Akaike Information Criterion (AIC) [92] or the Bayesian Information Criterion (BIC) [93] allow to quantify the improvement in the fit and penalize model complexity, and are computationally less demanding than Bayesian approaches like the Bayes factor ratio [94]. However, these criteria usually rely on differences in χ^2 at the maximum a posteriori (MAP) points, which may be misleading when applied to synthetic data sets such as our mock 3×2pt observation. Hence, we instead estimate the preference as the Mahalanobis distance in the 2-dimensional w_0-w_a parameter space, which serves as a meaningful metric given that our posteriors are nearly Gaussian. In particular, we compute

$$Q_{\boldsymbol{\theta}} = \sqrt{(\vec{w}_{\rm Fid} - \vec{w}_\Lambda)^T \boldsymbol{\Sigma}_{w_0 w_a}^{-1} (\vec{w}_{\rm Fid} - \vec{w}_\Lambda)}, \tag{D1}$$

where $\vec{w}_{\rm Fid} = (w_0, w_a) = (-0.8276, -0.6653)$ is our fiducial point, $\vec{w}_\Lambda = (w_0, w_a) = (-1, 0)$ is the ΛCDM limit, and $\boldsymbol{\Sigma}_{w_0 w_a}$ is the parameter covariance matrix reconstructed from the analysis. The metric $Q_{\boldsymbol{\theta}}$ in Eq. D1 generalizes the "rule of thumb difference in mean" [95] to various parameters, while incorporating the correlations between them. In Tab. III we report the values of $Q_{\boldsymbol{\theta}}$ for the different data sets (Baseline, 3×2pt, Baseline+3×2pt) and methods (MCMC, SBI) employed in the $w_0 w_a$CDM analysis. We find excellent agreement between the $Q_{\boldsymbol{\theta}}$ values obtained using MCMC and SBI, which was expected given the good overlap of their corresponding contours shown in Sec. IV.

[1] N. Aghanim *et al.* (Planck), Planck 2018 results. VI. Cosmological parameters, Astron. Astrophys. **641**, A6 (2020), [Erratum: Astron.Astrophys. 652, C4 (2021)], arXiv:1807.06209 [astro-ph.CO].

[2] T. Louis *et al.* (ACT), The Atacama Cosmology Telescope: DR6 Power Spectra, Likelihoods and ΛCDM Parameters, (2025), arXiv:2503.14452 [astro-ph.CO].

[3] E. Camphuis *et al.* (SPT-3G), SPT-3G D1: CMB temperature and polarization power spectra and cosmology from 2019 and 2020 observations of the SPT-3G Main field, (2025), arXiv:2506.20707 [astro-ph.CO].

[4] A. H. Wright *et al.*, KiDS-Legacy: Cosmological constraints from cosmic shear with the complete Kilo-Degree Survey, (2025), arXiv:2503.19441 [astro-ph.CO].

[5] T. M. C. Abbott *et al.* (DES), Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing, Phys. Rev. D **105**, 023520 (2022), arXiv:2105.13549 [astro-ph.CO].

[6] S. Alam *et al.* (eBOSS), Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory, Phys. Rev. D **103**, 083533 (2021), arXiv:2007.08991 [astro-ph.CO].

[7] M. Abdul Karim *et al.* (DESI), DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints, (2025), arXiv:2503.14738 [astro-ph.CO].

[8] Y. Mellier *et al.* (Euclid), Euclid. I. Overview of the Euclid mission, (2024), arXiv:2405.13491 [astro-ph.CO].

[9] Ž. Ivezić *et al.* (LSST), LSST: from Science Drivers to Reference Design and Anticipated Data Products, Astrophys. J. **873**, 111 (2019), arXiv:0805.2366 [astro-ph].

[10] P. Ade *et al.* (Simons Observatory), The Simons Observatory: Science goals and forecasts, JCAP **02**, 056, arXiv:1808.07445 [astro-ph.CO].

[11] K. N. Abazajian *et al.* (CMB-S4), CMB-S4 Science Book, First Edition, (2016), arXiv:1610.02743 [astro-ph.CO].

[12] A. Lewis, A. Challinor, and A. Lasenby, Efficient computation of CMB anisotropies in closed FRW models, Astrophys. J. **538**, 473 (2000), arXiv:astro-ph/9911177.

[13] J. Lesgourgues, The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview, (2011), arXiv:1104.2932 [astro-ph.IM].

[14] N. Schöneberg, G. Franco Abellán, A. Pérez Sánchez, S. J. Witte, V. Poulin, and J. Lesgourgues, The H0 Olympics: A fair ranking of proposed models, Phys. Rept. **984**, 1 (2022), arXiv:2107.10291 [astro-ph.CO].

[15] T. Brinckmann, D. C. Hooper, M. Archidiacono, J. Lesgourgues, and T. Sprenger, The promising future of a robust cosmological neutrino mass measurement, JCAP **01**, 059, arXiv:1808.05955 [astro-ph.CO].

[16] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, Proc. Nat. Acad. Sci. **117**, 30055 (2020), arXiv:1911.01429 [stat.ML].

[17] P. Lemos *et al.* (SimBIG), Field-level simulation-based inference of galaxy clustering with convolutional neural networks, Phys. Rev. D **109**, 083536 (2024), arXiv:2310.15256 [astro-ph.CO].

[18] N. Jeffrey *et al.* (DES), Dark energy survey year 3 results: likelihood-free, simulation-based wCDM inference with neural compression of weak-lensing map statistics, Mon. Not. Roy. Astron. Soc. **536**, 1303 (2024), arXiv:2403.02314 [astro-ph.CO].

[19] 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, Astron. Astrophys. **694**, A223 (2025), arXiv:2404.15402 [astro-ph.CO].

[20] C. P. Novaes, L. Thiele, J. Armijo, S. Cheng, J. A. Cowell, G. A. Marques, E. G. M. Ferreira, M. Shirasaki, K. Osato, and J. Liu, Cosmology from HSC Y1 weak lensing data with combined higher-order statistics and simulation-based inference, Phys. Rev. D **111**, 083510 (2025), arXiv:2409.01301 [astro-ph.CO].

[21] J. Zeghal, D. Lanzieri, F. Lanusse, A. Boucaud, G. Louppe, E. Aubourg, and A. E. Bayer (LSST Dark Energy Science), Simulation-Based Inference Benchmark for Weak Lensing Cosmology, (2024), arXiv:2409.17975 [astro-ph.CO].

[22] 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?, Phys. Rev. Lett. **133**, 221006 (2024), arXiv:2403.03220 [astro-ph.CO].

[23] Í. Zubeldia, B. Bolliet, A. Challinor, and W. Handley, Extracting cosmological information from the abundance of galaxy clusters with simulation-based inference, (2025), arXiv:2504.10230 [astro-ph.CO].

[24] G. Franco Abellán, G. C. n. Herrera, M. Martinelli, O. Savchenko, D. Sciotti, and C. Weniger, Fast likelihood-free inference in the LSS Stage IV era, JCAP **11**, 057, arXiv:2403.14750 [astro-ph.CO].

[25] O. Savchenko, G. Franco Abellán, F. List, N. Anau Montel, and C. Weniger, Fast Sampling of Cosmological Initial Conditions with Gaussian Neural Posterior Estimation, (2025), arXiv:2502.03139 [astro-ph.CO].

[26] N. Aghanim *et al.* (Planck), Planck 2018 results. V. CMB power spectra and likelihoods, Astron. Astrophys. **641**, A5 (2020), arXiv:1907.12875 [astro-ph.CO].

[27] 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, Mon. Not. Roy. Astron. Soc. **511**, 1771 (2022), arXiv:2106.03846 [astro-ph.CO].

[28] A. Nygaard, E. B. Holm, S. Hannestad, and T. Tram, CONNECT: a neural network based framework for emulating cosmological observables and cosmological parameter inference, JCAP **05**, 025, arXiv:2205.15726 [astro-ph.IM].

[29] S. Günther, L. Balkenhol, C. Fidler, A. R. Khalife, J. Lesgourgues, M. R. Mosbech, and R. K. Sharma, OLÉ– Online Learning Emulation in Cosmology, (2025), arXiv:2503.13183 [astro-ph.CO].

[30] T. McClintock and E. Rozo, Reconstructing Probability Distributions with Gaussian Processes, Mon. Not. Roy. Astron. Soc. **489**, 4155 (2019), arXiv:1905.09299 [astro-ph.CO].

[31] H. T. J. Bevins, W. J. Handley, P. Lemos, P. H. Sims, E. d. L. Acedo, A. Fialkov, and J. Alsing, Marginal post-processing of Bayesian inference products with normalizing flows and kernel density estimators, Mon. Not. Roy. Astron. Soc. **526**, 4613 (2023), arXiv:2205.12841 [astro-ph.IM].

[32] A. Mootoovaloo, C. García-García, D. Alonso, and J. Ruiz-Zapatero, emuflow: normalizing flows for joint cosmological analysis, Mon. Not. Roy. Astron. Soc. **536**, 190 (2024), arXiv:2409.01407 [astro-ph.CO].

[33] P. L. Taylor, A. Cuceu, C.-H. To, and E. A. Zaborowski, CombineHarvesterFlow: Joint Probe Analysis Made Easy with Normalizing Flows 10.33232/001c.124495 (2024), arXiv:2406.06687 [astro-ph.CO].

[34] M. Chevallier and D. Polarski, Accelerating universes with scaling dark matter, Int. J. Mod. Phys. D **10**, 213 (2001), arXiv:gr-qc/0009008.

[35] E. V. Linder, Exploring the expansion history of the universe, Phys. Rev. Lett. **90**, 091301 (2003), arXiv:astro-ph/0208512.

[36] A. G. Adame *et al.* (DESI), DESI 2024 VI: cosmological constraints from the measurements of baryon acoustic oscillations, JCAP **02**, 021, arXiv:2404.03002 [astro-ph.CO].

[37] A. Lewis and E. Chamberlain, Understanding acoustic scale observations: the one-sided fight against Λ, JCAP **05**, 065, arXiv:2412.13894 [astro-ph.CO].

[38] S. Nesseris, Y. Akrami, and G. D. Starkman, To CPL, or not to CPL? What we have not learned about the dark energy equation of state, (2025), arXiv:2503.22529 [astro-ph.CO].

[39] G. Efstathiou, Baryon Acoustic Oscillations from a Different Angle, (2025), arXiv:2505.02658 [astro-ph.CO].

[40] W. Giarè, Dynamical dark energy beyond Planck? Constraints from multiple CMB probes, DESI BAO, and type-Ia supernovae, Phys. Rev. D **112**, 023508 (2025), arXiv:2409.17074 [astro-ph.CO].

[41] F. B. M. d. Santos, J. Morais, S. Pan, W. Yang, and E. Di Valentino, A New Window on Dynamical Dark Energy: Combining DESI-DR2 BAO with future Gravitational Wave Observations, (2025), arXiv:2504.04646 [astro-ph.CO].

[42] K. Lodha *et al.* (DESI), Extended Dark Energy analysis using DESI DR2 BAO measurements, (2025), arXiv:2503.14743 [astro-ph.CO].

[43] G. Ye, M. Martinelli, B. Hu, and A. Silvestri, Hints of Nonminimally Coupled Gravity in DESI 2024 Baryon Acoustic Oscillation Measurements, Phys. Rev. Lett. **134**, 181002 (2025), arXiv:2407.15832 [astro-ph.CO].

[44] T.-N. Li, Y.-H. Li, G.-H. Du, P.-J. Wu, L. Feng, J.-F. Zhang, and X. Zhang, Revisiting holographic dark energy after DESI 2024, Eur. Phys. J. C **85**, 608 (2025), arXiv:2411.08639 [astro-ph.CO].

[45] W. Elbers, C. S. Frenk, A. Jenkins, B. Li, and S. Pascoli, Negative neutrino masses as a mirage of dark energy, Phys. Rev. D **111**, 063534 (2025), arXiv:2407.10965 [astro-ph.CO].

[46] W. J. Wolf, C. García-García, and P. G. Ferreira, Robustness of dark energy phenomenology across different parameterizations, JCAP **05**, 034, arXiv:2502.04929 [astro-ph.CO].

[47] S. Nakagawa, Y. Nakai, Y.-C. Qiu, and M. Yamada, Interpreting Cosmic Birefringence and DESI Data with Evolving Axion in ΛCDM, (2025), arXiv:2503.18924 [astro-ph.CO].

[48] L. A. Ureña-López *et al.*, Updated cosmological constraints on axion dark energy with DESI, (2025), arXiv:2503.20178 [astro-ph.CO].

[49] S.-F. Chen and M. Zaldarriaga, It's All `Ok`: Curvature in Light of BAO from DESI DR2, (2025), arXiv:2505.00659 [astro-ph.CO].

[50] D. Camarena, K. Greene, J. Houghteling, and F.-Y. Cyr-Racine, DESIgning concordant distances in the age of precision cosmology: the impact of density fluctuations, (2025), arXiv:2507.17969 [astro-ph.CO].

[51] A. Blanchard *et al.* (Euclid), Euclid preparation. VII. Forecast validation for Euclid cosmological probes, Astron. Astrophys. **642**, A191 (2020), arXiv:1910.09273

[astro-ph.CO].

[52] G. Papamakarios and I. Murray, Fast ϵ-free inference of simulation models with bayesian conditional density estimation, in *Advances in Neural Information Processing Systems* (2016) pp. 1028–1036.

[53] N. Jeffrey and B. D. Wandelt, Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks, in *34th Conference on Neural Information Processing Systems* (2020) arXiv:2011.05991 [stat.ML].

[54] G. Papamakarios, D. Sterratt, and I. Murray, Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, in *Proceedings of he 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019) pp. 837–848.

[55] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, Fast likelihood-free cosmology with neural density estimators and active learning, Mon. Not. Roy. Astron. Soc. **488**, 4440 (2019), arXiv:1903.00007 [astro-ph.CO].

[56] K. Cranmer, J. Pavez, and G. Louppe, Approximating Likelihood Ratios with Calibrated Discriminative Classifiers, (2015), arXiv:1506.02169 [stat.AP].

[57] J. Hermans, V. Begy, and G. Louppe, Likelihood-free MCMC with amortized approximate ratio estimators, in *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 (PMLR, 2020) pp. 4239–4248.

[58] M. Archidiacono *et al.* (Euclid), Euclid preparation - LIV. Sensitivity to neutrino parameters, Astron. Astrophys. **693**, A58 (2025), arXiv:2405.06047 [astro-ph.CO].

[59] R. Kou and A. Lewis, Improving CMB constraints on early Universe physics with LSS: A multi-probe forecast including cross-covariance, (2025), arXiv:2504.13913 [astro-ph.CO].

[60] W. Hu and I. Sawicki, A Parameterized Post-Friedmann Framework for Modified Gravity, Phys. Rev. D **76**, 104043 (2007), arXiv:0708.1190 [astro-ph].

[61] W. Fang, W. Hu, and A. Lewis, Crossing the Phantom Divide with Parameterized Post-Friedmann Dark Energy, Phys. Rev. D **78**, 087303 (2008), arXiv:0808.3125 [astro-ph].

[62] D. Blas, J. Lesgourgues, and T. Tram, The Cosmic Linear Anisotropy Solving System (CLASS) II: Approximation schemes, JCAP **07**, 034, arXiv:1104.2933 [astro-ph.CO].

[63] R. E. Smith, J. A. Peacock, A. Jenkins, S. D. M. White, C. S. Frenk, F. R. Pearce, P. A. Thomas, G. Efstathiou, and H. M. P. Couchmann (VIRGO Consortium), Stable clustering, the halo model and nonlinear cosmological power spectra, Mon. Not. Roy. Astron. Soc. **341**, 1311 (2003), arXiv:astro-ph/0207664.

[64] R. Takahashi, M. Sato, T. Nishimichi, A. Taruya, and M. Oguri, Revising the Halofit Model for the Nonlinear Matter Power Spectrum, Astrophys. J. **761**, 152 (2012), arXiv:1208.2701 [astro-ph.CO].

[65] N. Aghanim *et al.* (Planck), Planck 2018 results. VIII. Gravitational lensing, Astron. Astrophys. **641**, A8 (2020), arXiv:1807.06210 [astro-ph.CO].

[66] D. Scolnic *et al.*, The Pantheon+ Analysis: The Full Data Set and Light-curve Release, Astrophys. J. **938**, 113 (2022), arXiv:2112.03863 [astro-ph.CO].

[67] D. Brout *et al.*, The Pantheon+ Analysis: Cosmological Constraints, Astrophys. J. **938**, 110 (2022), arXiv:2202.04077 [astro-ph.CO].

[68] M. Martinelli *et al.* (Euclid), Euclid: Impact of nonlinear and baryonic feedback prescriptions on cosmological parameter estimation from weak lensing cosmic shear, Astron. Astrophys. **649**, A100 (2021), arXiv:2010.12382 [astro-ph.CO].

[69] S. Ilić *et al.* (Euclid), Euclid preparation. XV. Forecasting cosmological constraints for the Euclid and CMB joint analysis, Astron. Astrophys. **657**, A91 (2022), arXiv:2106.08346 [astro-ph.CO].

[70] B. Audren, J. Lesgourgues, K. Benabed, and S. Prunet, Conservative Constraints on Early Cosmology: an illustration of the Monte Python cosmological parameter inference code, JCAP **02**, 001, arXiv:1210.7183 [astro-ph.CO].

[71] T. Brinckmann and J. Lesgourgues, MontePython 3: boosted MCMC sampler and other features, Phys. Dark Univ. **24**, 100260 (2019), arXiv:1804.07261 [astro-ph.CO].

[72] A. Gelman and D. B. Rubin, Inference from Iterative Simulation Using Multiple Sequences, Statist. Sci. **7**, 457 (1992).

[73] A. Lewis, GetDist: a Python package for analysing Monte Carlo samples, (2019), arXiv:1910.13970 [astro-ph.IM].

[74] B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, Truncated Marginal Neural Ratio Estimation, in *35th Conference on Neural Information Processing Systems* (2021) arXiv:2107.01214 [stat.ML].

[75] A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino, and C. Weniger, Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation, JCAP **09**, 004, arXiv:2111.08030 [astro-ph.CO].

[76] N. A. Montel, A. Coogan, C. Correa, K. Karchev, and C. Weniger, Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation, Mon. Not. Roy. Astron. Soc. **518**, 2746 (2022), arXiv:2205.09126 [astro-ph.CO].

[77] K. Karchev, R. Trotta, and C. Weniger, SICRET: Supernova Ia Cosmology with truncated marginal neural Ratio EsTimation, Mon. Not. Roy. Astron. Soc. **520**, 1056 (2023), arXiv:2209.06733 [astro-ph.CO].

[78] A. Saxena, A. Cole, S. Gazagnes, P. D. Meerburg, C. Weniger, and S. J. Witte, Constraining the X-ray heating and reionization using 21-cm power spectra with Marginal Neural Ratio Estimation, Mon. Not. Roy. Astron. Soc. **525**, 6097 (2023), arXiv:2303.07339 [astro-ph.CO].

[79] J. Alvey, M. Gerdes, and C. Weniger, Albatross: a scalable simulation-based inference pipeline for analysing stellar streams in the Milky Way, Mon. Not. Roy. Astron. Soc. **525**, 3662 (2023), arXiv:2304.02032 [astro-ph.GA].

[80] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, Sequential simulation-based inference for gravitational wave signals, Phys. Rev. D **108**, 042004 (2023), arXiv:2304.02035 [gr-qc].

[81] P. S. Cole, J. Alvey, L. Speri, C. Weniger, U. Bhardwaj, D. Gerosa, and G. Bertone, Sequential simulation-based inference for extreme mass ratio inspirals, (2025), arXiv:2505.16795 [gr-qc].

[82] V. Chiley, I. Sharapov, A. Kosson, U. Koster, R. Reece, S. S. de la Fuente, V. Subbiah, and M. James, Online normalization for training neural networks, (2019), arXiv:1905.05894 [cs.LG].

[83] E. Rosenberg, S. Gratton, and G. Efstathiou, CMB power spectra and cosmological parameters from Planck PR4 with CamSpec, Mon. Not. Roy. Astron. Soc. **517**,

4620 (2022), arXiv:2205.10869 [astro-ph.CO].

[84] M. S. Madhavacheril *et al.* (ACT), The Atacama Cosmology Telescope: DR6 Gravitational Lensing Map and Cosmological Parameters, Astrophys. J. **962**, 113 (2024), arXiv:2304.05203 [astro-ph.CO].

[85] A. Spurio Mancini, K. Lin, and J. D. McEwen, Field-level cosmological model selection: field-level simulation-based inference for Stage IV cosmic shear can distinguish dynamical dark energy, (2024), arXiv:2410.10616 [astro-ph.CO].

[86] G. E. Addison, T. M. Essinger-Hileman, M. R. Greason, T. B. Griswold, T. Jaffe, N. Miller, U. Prasad, and J. L. Weiland, Legacy Archive for Microwave Background Data Analysis (LAMBDA): An Overview, (2019), arXiv:1905.08667 [astro-ph.IM].

[87] E. Di Valentino *et al.* (CosmoVerse), The CosmoVerse White Paper: Addressing observational tensions in cosmology with systematics and fundamental physics 10.1016/j.dark.2025.101965 (2025), arXiv:2504.01669 [astro-ph.CO].

[88] 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).

[89] G. Papamakarios, T. Pavlakou, and I. Murray, Masked Autoregressive Flow for Density Estimation, (2017), arXiv:1705.07057 [stat.ML].

[90] J. U. Lange, nautilus: boosting Bayesian importance nested sampling with deep learning, Mon. Not. Roy. Astron. Soc. **525**, 3181 (2023), arXiv:2306.16923 [astro-ph.IM].

[91] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, V. Begy, and G. Louppe, A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful, (2021), arXiv:2110.06581 [stat.ML].

[92] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automatic Control **19**, 716 (1974).

[93] G. Schwarz, Estimating the Dimension of a Model, Annals Statist. **6**, 461 (1978).

[94] R. E. Kass and A. E. Raftery, Bayes Factors, J. Am. Statist. Assoc. **90**, 773 (1995).

[95] M. Raveri and W. Hu, Concordance and Discordance in Cosmology, Phys. Rev. D **99**, 043506 (2019), arXiv:1806.04649 [astro-ph.CO].