

Impact of non-Gaussian likelihood on cosmological constraints from the thermal Sunyaev-Zel'dovich power spectrum: A simulation-based inference analysis

Licong Xu^{1,2,*}, Íñigo Zubeldia^{1,2,3}, James Alvey^{1,2}, Boris Bolliet^{4,2} and Anthony Challinor^{1,2,3}

¹*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*

²*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*

³*DAMTP, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom*

⁴*Cavendish Astrophysics, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*



(Received 12 June 2026; accepted 14 July 2026; published 14 August 2026)

The thermal Sunyaev–Zel'dovich (tSZ) power spectrum is a sensitive probe of cosmology and cluster astrophysics, but its statistics are non-Gaussian because the signal receives a significant contribution from rare, massive, low-redshift galaxy clusters. As a result, a Gaussian likelihood fails to describe the statistics of its power spectrum on large scales. We use simulation-based inference (SBI) to test the accuracy of the standard Gaussian power-spectrum likelihood for a *Planck*-like tSZ analysis. Using halo-based simulations of full-sky Compton- y maps, we train neural posterior and likelihood estimators and compare the resulting constraints with those from a Gaussian likelihood assumption. Using only multipoles $\ell < 1000$, we find that the Gaussian likelihood assumption gives unbiased cosmological constraints, while the SBI-based inference shows a mild broadening of the posterior distributions for the amplitudes of residual foregrounds. This suggests that the Gaussian likelihood assumption is sufficiently accurate for cosmological inference for a *Planck*-like tSZ analysis, while SBI provides a useful validation tool to model non-Gaussian likelihoods beyond analytic approximations.

DOI: [10.1103/tl9p-ldwf](https://doi.org/10.1103/tl9p-ldwf)

I. INTRODUCTION

Galaxy clusters are the largest gravitationally bound structures in the Universe. They form in the most massive dark-matter halos, with typical masses of a few $10^{14} M_{\odot}$. The abundance of galaxy clusters is highly sensitive to the cosmological model and the growth of cosmic structure across time. This has motivated numerous cosmological analyses using cluster number counts. In the past, the properties of galaxy clusters have been determined through different observations at different wavelengths, ranging from X-ray [see, e.g., [1] for a review] and optical [2–4] to Sunyaev–Zel'dovich (SZ) surveys [5–8].

The thermal Sunyaev–Zel'dovich (tSZ) effect [9] is one of the most important probes to study galaxy clusters. It produces a spectral distortion of the cosmic microwave

background (CMB) due to the inverse Compton scattering of the CMB photons by the hot electrons in the intracluster medium. Detecting such distortions allows one to construct tSZ maps tracing the large-scale distribution of hot gas. Cosmological and astrophysical information can then be extracted from these maps using a range of summary statistics (see, e.g., [10] for a review), and using cluster number counts derived from the tSZ signal (see, e.g., [11–14]).

Among all the summary statistics that can be extracted from tSZ maps, the power spectrum is one of the simplest and most direct probes of cosmological information. On large scales, the tSZ power spectrum depends strongly on the matter clustering parameter σ_8 , and matter density fraction Ω_m [15,16], and on the nonlinear gas physics within the clusters on small scales [17]. In particular, the tSZ power spectrum contains information about all the clusters that are not detected individually in cluster number-count analyses, and thus offers a complementary probe to the cluster number-count method [18].

Previous analyses of the tSZ power spectrum have typically assumed a Gaussian likelihood for the power spectrum. For example, the Planck Collaboration [19]

*Contact author: lx256@cam.ac.uk

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

analysis constrained both the cosmological parameters and residual foreground amplitudes, using a covariance based only on the Gaussian contribution. Bolliet *et al.* [20] revisited this approach by incorporating the non-Gaussian trispectrum contribution to the covariance, and by varying all relevant cosmological and foreground parameters simultaneously. More recent analyses include a *Planck* tSZ power spectrum analysis based on the PR4 data [21], and a measurement of the full shape of the tSZ power spectrum from the South Pole Telescope [22]. However, the contribution to the tSZ power spectrum is dominated by massive clusters, whose rarity introduces non-Gaussianity. This effect is more significant on large angular scales, where the signal is dominated by a small number of the largest clusters, and is less prominent on smaller scales as the contribution becomes distributed over a large number of less-massive clusters. Several studies (e.g., [23,24]) have shown that the sampling distribution of the tSZ power spectrum is non-Gaussian and positively skewed at low multipoles even for large sky fractions, approaching Gaussianity only at higher multipoles and for large sky fractions. Importantly, this type of non-Gaussianity is beyond that due to the finite-mode statistics of the power spectrum estimator at low multipoles. Nevertheless, a likelihood model that accurately captures the non-Gaussian statistics of the tSZ power spectrum remains unavailable.

In this work, we propose an alternative method to obtain cosmological constraints from the tSZ power spectrum using a simulation-based inference (SBI) approach (see, e.g., [25] for a comprehensive review). SBI enables parameter inference directly from forward-modeled simulations, without requiring an explicit analytic likelihood. Working in a *Planck*-like setting, we generate mock galaxy cluster catalogs and produce their corresponding sky maps for a given set of cosmological parameters, including residual foreground contamination. We then train deep neural networks either to learn the posterior distribution directly via Neural Posterior Estimation (NPE; [26]) or to learn the likelihood function through Neural Likelihood Estimation (NLE; [27]). We validate our SBI-based approach using mock sky maps generated across a range of input parameters with the same pipeline as that used for training, and compare the consistency between the learned likelihood and the true sampling distribution.

This paper is organized as follows. In Secs. II and III, we review the theoretical calculation of the tSZ power spectrum within the halo-model formalism and its statistical properties. Next, in Sec. IV, we outline the Gaussian-likelihood formalism of the tSZ power spectrum for constraining parameters. Section V describes our forward-modeling approaches for the tSZ power spectrum, including both a Gaussian-likelihood-based simulator and the halo-based simulations that generate full-sky maps. In Sec. VI, we present our SBI-based approach and the details

of the neural density estimation implementations. The main parameter results are given in Sec. VII, and we validate our SBI-based approach in Sec. VIII. We conclude with a discussion and outlook in Sec. IX.

II. THEORY OF THE TSZ POWER SPECTRUM

A. Halo-model formalism of the tSZ power spectrum

The tSZ effect introduces a frequency-dependent shift of the observed CMB temperature. In the nonrelativistic limit, it can be written as [9]

$$\frac{\Delta T(\nu)}{T_{\text{CMB}}} = f(\nu)y(\theta), \quad (1)$$

where $\Delta T(\nu)$ is the frequency-dependent temperature deviation, $T_{\text{CMB}} = 2.7255$ K is the temperature of the undistorted CMB [28], $f(\nu) = x \coth(x/2) - 4$ is the tSZ spectral function with $x = h_P \nu / (k_B T_{\text{CMB}})$ and h_P , k_B are the Planck and Boltzmann constants, respectively. Also, $y(\theta)$ is known as the Compton- y parameter, which is the line-of-sight integral of electron pressure as a function of angular position θ :

$$y(\theta) = \frac{\sigma_T}{m_e c^2} \int P_e \left(\sqrt{l^2 + d_A^2 \theta^2} \right) dl, \quad (2)$$

where σ_T is the Thomson cross section, m_e is the electron mass, c is the speed of light, P_e is the electron pressure, l is the line-of-sight distance, and d_A is the angular diameter distance.

We adopt a halo-model formalism [29] to model the tSZ power spectrum. In this formalism, the tSZ power spectrum can be written as the sum of 1-halo and 2-halo terms. The 1-halo term is the sum of the squared two-dimensional Fourier transform of the pressure profile from individual clusters. In the continuous limit, the 1-halo term can be written as

$$C_\ell^{\text{tSZ,1h}} = \int_0^{z_{\text{max}}} dz \frac{dV}{dz d\Omega} \int_{M_{\text{min}}}^{M_{\text{max}}} dM \frac{dn(M, z)}{dM} |\tilde{y}_\ell(M, z)|^2, \quad (3)$$

where $dV/(dz d\Omega)$ is the derivative of comoving volume V with respect to redshift z and solid angle Ω , and dn/dM is the halo mass function (HMF), for which we adopt the form of Tinker *et al.* [30]. The quantity $\tilde{y}_\ell$ is the two-dimensional Fourier transform of the (spherically symmetric) Compton- y profile, $y(\theta)$, for a cluster of mass M and redshift z given by

$$\tilde{y}_\ell(M, z) = \frac{4\pi r_s}{\ell_s^2} \frac{\sigma_T}{m_e c^2} \int_0^\infty dx x^2 \frac{\sin(\ell x / \ell_s)}{\ell x / \ell_s} P_e(x; M, z), \quad (4)$$

where $x = r/r_s$, with r being the radial distance to the center of the halo and r_s the characteristic scale radius, and $\ell_s = d_A/r_s$. In this paper, we set $r_s = r_{500c}/c_{500}$, where r_{500c} is the radius within which the mean enclosed density is 500 times the critical density of the universe at the cluster redshift, and c_{500} is the corresponding concentration parameter. The 2-halo term describes the correlation between spatial positions of the clusters, and is relevant on large angular scales. The 2-halo term is not significant compared to the 1-halo term when all the massive and local clusters are included, but can have a non-negligible contribution when those clusters are masked out [15]. We neglect the 2-halo term in this paper since we include the rare massive and local clusters in our analysis.

The electron pressure profile is assumed to follow the generalized Navarro–Frenk–White (GNFW) profile given by

$$P_e(x) = C \times P_0(c_{500}x)^{-\gamma} [1 + (c_{500}x)^\alpha]^{(\gamma-\beta)/\alpha}, \quad (5)$$

where, recall, $x = r/r_s$. In this paper, we adopt the pressure profile of Arnaud *et al.* [31], hereafter A10, where the parameters $\{\gamma, \alpha, \beta, P_0, c_{500}\}$ are set to be constants, and the coefficient C is given by

$$C = 1.65 \left(\frac{h}{0.7} \right)^2 \left(\frac{H(z)}{H_0} \right)^{8/3} \times \left(\frac{\tilde{M}_{500c}}{0.7 \times 3 \times 10^{14} M_\odot} \right)^{2/3+0.12} \text{ eV cm}^{-3}. \quad (6)$$

Here, $H(z)$ is the Hubble parameter, with present-day value $H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1}$. The mass $\tilde{M}_{500c}$ is related to the true mass via the relation $\tilde{M}_{500c} = M_{500c}/B$, where M_{500c} is the mass enclosed within a radius inside which the mean density is 500 times the critical density of the universe at redshift z , and the bias factor B accounts for the difference between X-ray mass proxies and the true halo mass that affects the scaling relations. In the literature, a different variable $1 - b$ is used more often and it is related to B via $B = 1/(1 - b)$. This can be caused by nonhydrostatic pressure support, such as turbulence, magnetic fields, or cosmic rays [32], instrumental calibration, or selection effects [33]. Since r_{500c} is defined via the mass estimates, a biased mass estimate $\tilde{M}_{500c} = M_{500c}/B$ in the pressure profile scaling relation also changes the corresponding scale radius, giving $r_{500c} \propto (M_{500c}/B)^{1/3}$.

Throughout the paper, we assume a flat Λ CDM cosmology with an effective number of neutrino species $N_{\text{eff}} = 3.046$, and the sum of neutrino masses to be $\sum m_\nu = 0.06 \text{ eV}$. Unless otherwise stated, all benchmark cluster catalogs and Compton- y maps are produced at fiducial cosmological parameters defined by $\omega_b = 0.02242$, $\omega_{\text{cdm}} = 0.1193$, $H_0 = 67.66 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\ln(10^{10} A_s) = 2.9718$, $n_s = 0.9665$, and a fiducial hydrostatic bias parameter $B = 1.41$, which we take to be independent of mass and redshift.

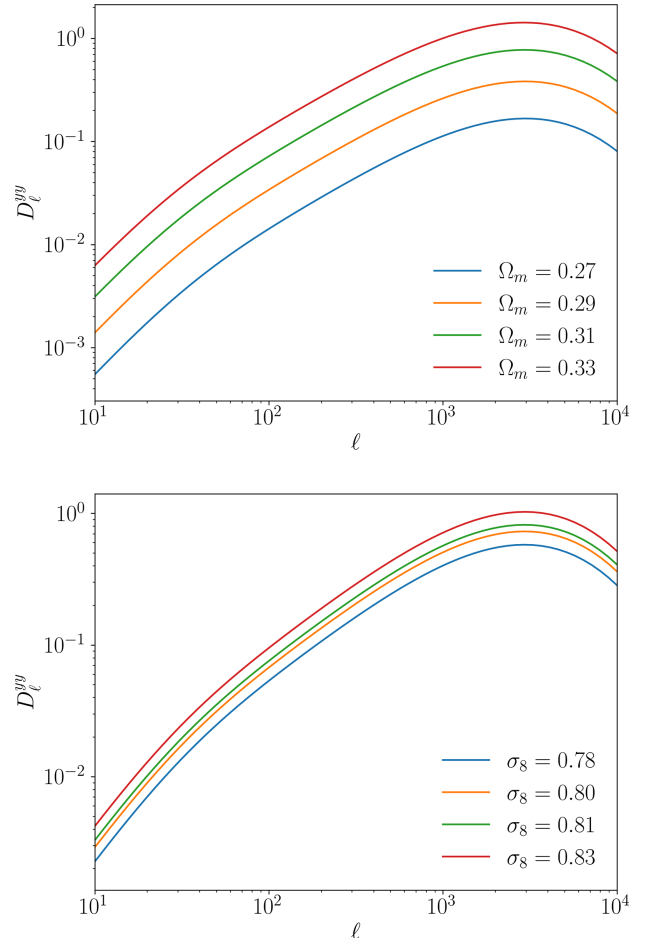


FIG. 1. Dependence of the tSZ power spectrum, $D_\ell^{yy} = 10^{12} \times \ell(\ell + 1) C_\ell^{\text{tSZ}} / (2\pi)$, on Ω_m (upper) and σ_8 (lower) with fixed ω_b , h and n_s .

B. Parameter dependence of the tSZ power spectrum

On large scales, the shape of the tSZ power spectrum is sensitive to the matter density, Ω_m , and amplitude of matter fluctuations, σ_8 . Figure 1 illustrates how variations in Ω_m and σ_8 affect the shape of the tSZ power spectrum. For the multipole range relevant to the *Planck* survey ($\ell < 10^3$), Komatsu and Seljak [16] find that the tSZ power spectrum scales approximately as

$$C_\ell^{\text{tSZ}} \propto \sigma_8^{8.1} \Omega_m^{3.2} B^{-3.2} h^{-1.7}. \quad (7)$$

Hence, the tSZ power spectrum is sensitive to the parameter combination F defined as $F = \sigma_8(\Omega_m/B)^{0.40} h^{-0.21}$.

C. Mass and redshift dependence of the tSZ power spectrum

We now investigate the halo masses and redshifts to which the tSZ power spectrum is most sensitive at $\ell \leq 10^3$. The differential contributions of mass and redshift to the 1-halo tSZ power spectrum follow from Eq. (3). We have [16]

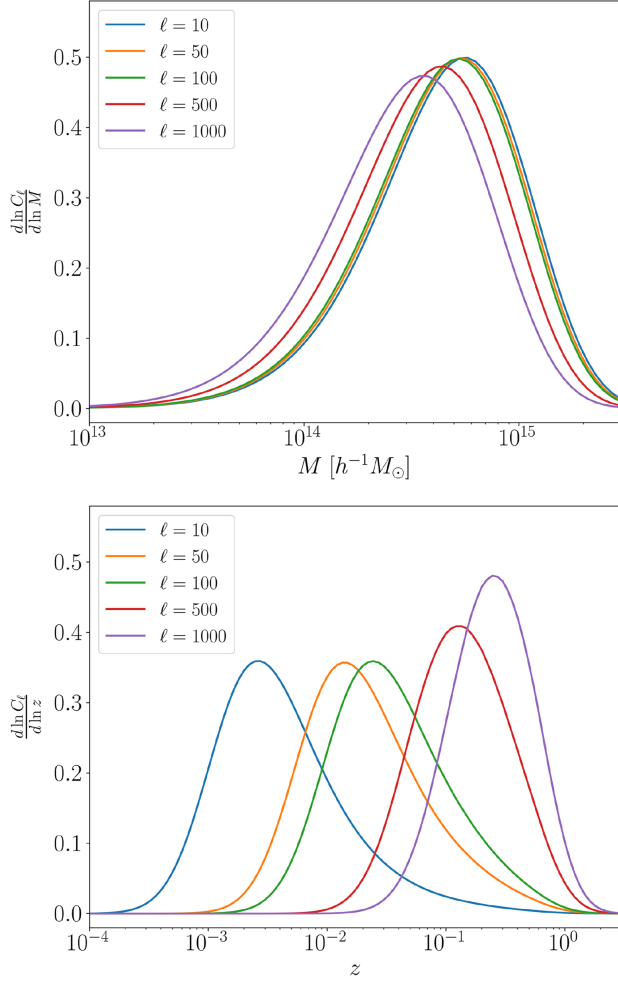


FIG. 2. Upper panel: Differential mass contribution to C_ℓ^{tSZ} , $d \ln C_\ell^{\text{tSZ}} / d \ln M$, for $\ell = 10, 50, 100, 500$, and 1000 . Lower panel: Differential redshift contribution to C_ℓ^{tSZ} , $d \ln C_\ell^{\text{tSZ}} / d \ln z$, for the same multipoles.

$$\frac{d \ln C_\ell^{\text{tSZ}}}{d \ln M} \equiv \frac{M \int dz \frac{dV}{dz d\Omega} \frac{dn(M, z)}{dM} |\tilde{y}_\ell(M, z)|^2}{\int dz \frac{dV}{dz d\Omega} \int dM \frac{dn(M, z)}{dM} |\tilde{y}_\ell(M, z)|^2}, \quad (8)$$

$$\frac{d \ln C_\ell^{\text{tSZ}}}{d \ln z} \equiv \frac{z \frac{dV}{dz d\Omega} \int dM \frac{dn(M, z)}{dM} |\tilde{y}_\ell(M, z)|^2}{\int dz \frac{dV}{dz d\Omega} \int dM \frac{dn(M, z)}{dM} |\tilde{y}_\ell(M, z)|^2}. \quad (9)$$

Figure 2 shows the differential mass and redshift distribution of C_ℓ^{tSZ} for multipoles $\ell = 10, 50, 100, 500$, and 1000 . We find that the contribution to the tSZ power spectrum gradually shifts toward lower-mass halos as ℓ increases, since higher multipoles probe smaller angular scales corresponding to more numerous small halos. At $\ell = 1000$, the mass distribution of C_ℓ^{tSZ} peaks at $M \sim 4 \times 10^{14} h^{-1} M_\odot$. It is the massive halos between $10^{14} h^{-1} M_\odot$ and $10^{15} h^{-1} M_\odot$ that dominate C_ℓ^{tSZ} over this multipole range of interest. At large angular scales ($\ell \lesssim 100$), the distribution converges, indicating that these modes are

already dominated by the most massive clusters. In this regime, further increasing the angular scale produces negligible changes in the range of halo masses contributing to the signal. In addition, the tSZ power spectrum is dominated by low-redshift halos for all these multipoles. The halos at $z \sim 10^{-2}$ dominate the contribution to C_ℓ^{tSZ} at $\ell \sim 100$, and halos at $z \sim 10^{-1}$ contribute to C_ℓ^{tSZ} at $\ell \sim 1000$. Halos with $z > 2$ have a negligible contribution to C_ℓ^{tSZ} for $\ell < 1000$, as their tSZ signal is strongly suppressed by the decrease in halo abundance and angular size with redshift.

III. STATISTICS OF THE TSZ POWER SPECTRUM

In this section, we discuss the sampling distribution of the tSZ power spectrum, both analytically and through simulations. We find that the distributions are significantly skewed due to rare, massive and local clusters, calling into question the usual Gaussian likelihood approximation.

We start by revisiting the derivation in Ref. [23] of the n -point joint probability distribution function (PDF) of the 1-halo term of the tSZ power spectrum. We extend it by including the scatter of the tSZ power spectrum given a fixed set of clusters on the sky due to their random locations.

Suppose there are N clusters contributing to the tSZ power spectrum, and we bin these into α bins (e.g., mass, redshift and solid angle). Let N_i be the number of clusters in the i th bin. The 1-point distribution of the tSZ power spectrum C at specific multipole ℓ , $P(C)$, can be written as

$$P(C) = \sum_{N_1=0}^{\infty} \dots \sum_{N_\alpha=0}^{\infty} P(N_1, \dots, N_\alpha) P(C|N_1, \dots, N_\alpha). \quad (10)$$

Assuming that clusters are drawn from the HMF according to a Poisson process, the number of clusters in each bin is independent of one another. The joint probability $P(N_1, \dots, N_\alpha)$ is then given by

$$P(N_1, \dots, N_\alpha) = \prod_{i=1}^{\alpha} \frac{\lambda_i^{N_i}}{N_i!} \exp(-\lambda_i), \quad (11)$$

where λ_i is the mean number of clusters in each bin i . We also assume that for a given set of clusters $\{N_i\}$, the tSZ power spectrum has some scatter around the mean, which is assumed to be Gaussian with variance σ_G^2 . In principle, σ_G^2 is a function of the number of clusters N_i as derived in the Appendix A. Hence, $P(C|N)$ can be written as

$$P(C|N) = \frac{1}{\sqrt{2\pi\sigma_G^2}} \exp \left[-\frac{(C - \bar{C})^2}{2\sigma_G^2} \right], \quad (12)$$

where the mean is given by the 1-halo contribution from individual clusters, $\bar{C} = \frac{1}{\Omega_{\text{sky}}} \sum_{i=1}^{\alpha} N_i S_i$, with $S_i = |\tilde{y}_\ell(M_i, z_i)|^2$ being the square of the 2D Fourier transform

of the Compton- γ profile of the clusters in the i th bin. Also, $\Omega_{\text{sky}} = 4\pi f_{\text{sky}}$ is the sky area in steradians and f_{sky} is the sky fraction.

To evaluate the PDF of C , $P(C)$, it is convenient to consider the characteristic function of $P(C)$,

$$\begin{aligned}\tilde{P}(\omega) &= \int P(C) \exp(-i\omega C) dC \\ &= \exp \left\{ \sum_{i=1}^{\alpha} \lambda_i [\exp(-i\omega S_i / \Omega_{\text{sky}}) - 1] \right\} \\ &\quad \times \exp \left(-\frac{\sigma_G^2 \omega^2}{2} \right).\end{aligned}\quad (13)$$

In the continuous limit where the bin size goes to zero, the characteristic function becomes

$$\ln \tilde{P}(\omega) = \int d\tilde{N} [\exp(-i\omega S / \Omega_{\text{sky}}) - 1] - \frac{\sigma_G^2 \omega^2}{2}, \quad (14)$$

where the first term characterizes the Poisson process of cluster number counts modulated by the Compton- γ profile of the clusters, and the second term accounts for the Gaussian scatter of the power spectrum.

Suppose we bin the clusters in bins of mass, redshift, and solid angle, (M, z, Ω) . Then, in our notation, Eq. (14) can be written as

$$\begin{aligned}\ln \tilde{P}(\omega) &= \Omega_{\text{sky}} \int dz \frac{dV}{dz d\Omega} \int dM \frac{dn}{dM} \\ &\quad \times \left[\exp \left(-i\omega \frac{|\tilde{y}_{\ell}(M, z)|^2}{\Omega_{\text{sky}}} \right) - 1 \right] - \frac{\sigma_G^2 \omega^2}{2}.\end{aligned}\quad (15)$$

The formalism above can be easily generalized to the n -point joint PDF of the tSZ power spectrum, $P(C_1, \dots, C_n)$, by considering the joint characteristic function,

$$\begin{aligned}\ln \tilde{P}(\omega_1, \dots, \omega_n) &= \Omega_{\text{sky}} \int dz \frac{dV}{dz d\Omega} \int dM \frac{dn}{dM} \\ &\quad \times \left\{ \exp \left[-i \sum_{i=1}^n \omega_i \frac{|\tilde{y}_{\ell_i}(M, z)|^2}{\Omega_{\text{sky}}} \right] - 1 \right\} \\ &\quad - \frac{1}{2} \sum_{i=1}^n \omega_i^2 \sigma_{i,G}^2.\end{aligned}\quad (16)$$

Here, $\tilde{y}_{\ell_i}(M, z)$ is the two-dimensional Fourier transform of the cluster Compton- γ profile at multipole ℓ_i in Eq. (16), and $\sigma_{i,G}^2$ denotes the Gaussian contribution to the variance in the corresponding multipole bin. The connected moments of the C_i follow from $\ln \tilde{P}(\omega_1, \dots, \omega_n)$ by the cumulant expansion theorem. Expanding the right-hand side of Eq. (16) in powers of ω , gives

$$\begin{aligned}&\exp \left[-i \sum_{i=1}^n \omega_i \frac{|\tilde{y}_{\ell_i}(M, z)|^2}{\Omega_{\text{sky}}} \right] - 1 \\ &= -i \sum_{i=1}^n \omega_i \frac{|\tilde{y}_{\ell_i}(M, z)|^2}{\Omega_{\text{sky}}} \\ &\quad - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \omega_i \omega_j \frac{|\tilde{y}_{\ell_i}(M, z)|^2 |\tilde{y}_{\ell_j}(M, z)|^2}{\Omega_{\text{sky}}^2} + \dots\end{aligned}\quad (17)$$

The linear term in ω generates the mean of $C_{\ell}^{\text{tSZ,1h}}$ given by Eq. (3), and the quadratic term yields the covariance $M_{\ell\ell'}$ of $C_{\ell}^{\text{tSZ,1h}}$ between different multipoles, given by

$$M_{\ell\ell'} = \sigma_{\ell,G}^2 \delta_{\ell\ell'} + \frac{T_{\ell\ell'}^{\text{yy}}}{4\pi f_{\text{sky}}}, \quad (18)$$

where the trispectrum $T_{\ell\ell'}^{\text{yy}}$ is given by

$$T_{\ell\ell'}^{\text{yy}} = \int dz \frac{dV}{dz d\Omega} \int d \ln M \frac{dn}{d \ln M} |\tilde{y}_{\ell}|^2 |\tilde{y}_{\ell'}|^2. \quad (19)$$

We work with binned bandpowers, whose covariance matrix is written as

$$M_{bb'} = \frac{2(C_{\ell_b}^{\text{tSZ,1h}})^2}{(2\ell_b + 1)f_{\text{sky}}\Delta\ell_b} \delta_{bb'} + \frac{\bar{T}_{bb'}^{\text{yy}}}{4\pi f_{\text{sky}}}, \quad (20)$$

where ℓ_b is the effective multipole of bin b , $\Delta\ell_b$ is width, and the bin-averaged trispectrum is

$$\bar{T}_{bb'}^{\text{yy}} = \frac{1}{|\mathcal{B}_b||\mathcal{B}_{b'}|} \sum_{\ell \in \mathcal{B}_b} \sum_{\ell' \in \mathcal{B}_{b'}} T_{\ell\ell'}^{\text{yy}}, \quad (21)$$

with $\mathcal{B}_b$ being the set of integer multipoles in bin b . Note that σ_G^2 denotes the variance of the tSZ power spectrum conditioned on a fixed number of clusters. We evaluate this conditional variance at the mean number of clusters, at which point $C_{\ell_b}^{\text{tSZ,1h}}$ corresponds to the mean (theoretical) power spectrum.

A. Comparison to empirical sampling distributions

To quantify the non-Gaussianity of the tSZ power spectrum, we generate 1000 realizations of full-sky tSZ maps using halo-based simulations (described in Sec. VB) at the fiducial cosmology. The power spectrum is computed from each realization, binned into 18 multipole bins. Figure 3 shows the distribution of these binned power spectra compared to the analytic prediction based on numerical evaluation of Eq. (15) for the 1-point PDF. The agreement is excellent, reproducing the highly skewed distributions at low multipoles. We also compare the empirical distribution of the 1000 realizations with a Gaussian approximation of the binned covariance matrix

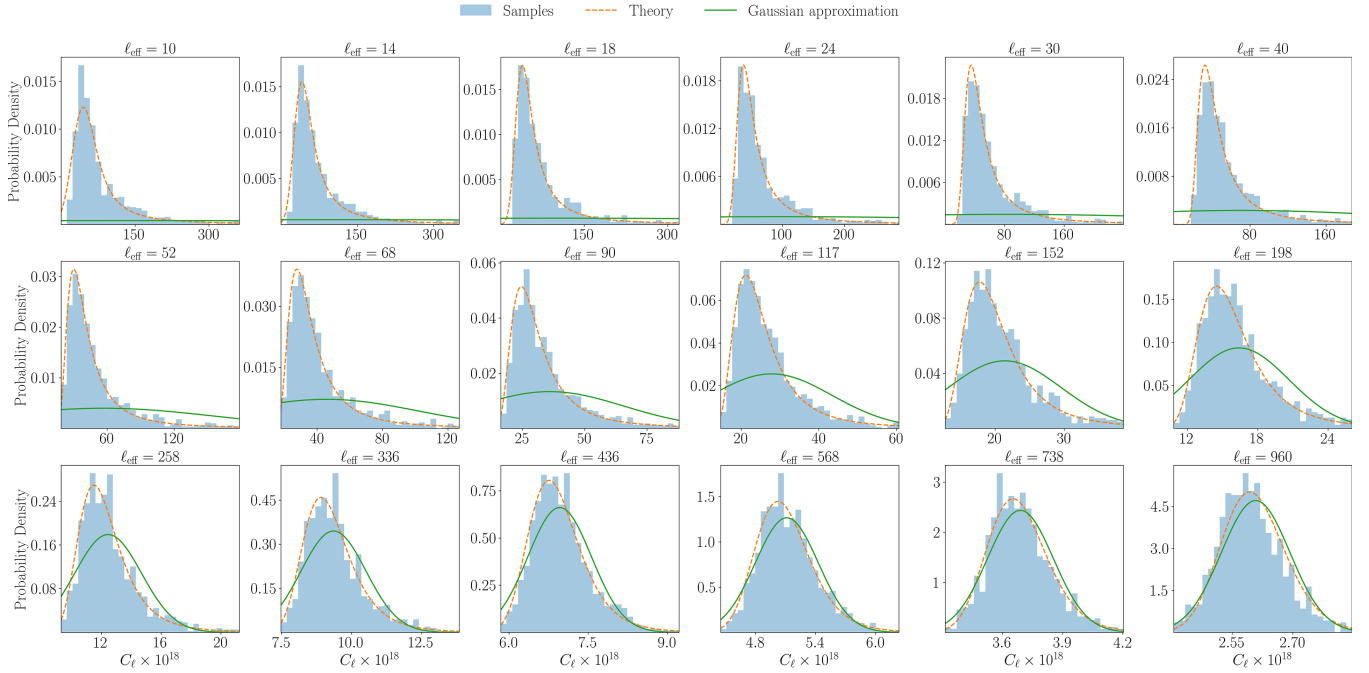


FIG. 3. One-point distribution of the binned tSZ power spectrum from 1000 realizations (blue) compared with the theoretical prediction and Gaussian approximation. The theoretical prediction (orange dashed) is computed from Eq. (15); the Gaussian approximation (green solid) uses a binned version of the covariance given by Eq. (20). The distribution obtained from simulations agrees well with the theoretical prediction at all multipoles.

$M_{bb'}$. As can be seen, the Gaussian approximation (green curve) fits the simulated data poorly at low multipoles, where the power spectrum is dominated by a small number of massive clusters, but improves at higher multipoles as the number of contributing clusters increases. The Gaussian prediction shown here includes both terms in Eq. (20), with the trispectrum dominating in this case. The trend toward Gaussian distributions at higher multipoles reflects the underlying Gaussian limit of the Poisson distribution from the central limit theorem, not an increasing contribution from the $\sigma_{\ell,G}$ term, which only captures the scatter in the power spectrum at fixed number of clusters.

For cosmological inference, we require the full n -point PDF since power spectra at different multipoles are not independent. The integral in Eq. (16) can be computed numerically for low-dimensional cases. However, for high-dimensional joint distributions, the dimensionality of the integral increases dramatically, and the full analytic evaluation of the n -point joint PDF becomes intractable. In such cases, SBI methods can provide an alternative route that does not require an explicit evaluation of the likelihood. SBI methods can estimate the underlying likelihood and perform parameter inference using simulations, as we explore in Secs. V and VI.

IV. GAUSSIAN LIKELIHOOD FORMALISM

Cosmological constraints can be obtained by modeling the observed tSZ power spectrum using Bayesian inference

methods. In real data, the total power spectrum of Compton- y maps includes the tSZ signal, in addition to residuals of astrophysical foregrounds and the instrumental noise. Here, we consider three astrophysical foreground components: the cosmic infrared background (CIB), shot noise from radio sources (RS) and from the infrared (IR) point sources that make up the CIB. These templates represent the foreground power that remains after applying the component-separation weights to multifrequency foreground models following the *Planck* y -map power-spectrum analysis [5]. A correlation between the tSZ signal and the CIB is expected because both trace the same large-scale structure: galaxy clusters host both the hot intracluster medium responsible for the tSZ effect and dusty star-forming galaxies that dominate the CIB. Following analyses by the Planck Collaboration [5], the shape of the tSZ and clustered CIB cross-power spectrum is very similar to that of the clustered CIB power spectrum [34]. Any cross-correlation power should therefore be absorbed by the CIB power spectrum template. Furthermore, since the tSZ power spectrum at multipoles $\ell \lesssim 1000$ is mostly sensitive to large clusters at redshifts $z \lesssim 1$, most of which are present in the *Planck* cluster catalogs, the analysis of Ref. [35] suggests that the cluster-correlated CIB on tSZ observables remains negligible for these large clusters. Turning to the IR and RS shot-noise templates, the shape of their power spectra is nearly white. In addition, following the *Planck* y -map power-spectrum analysis, we include an empirical correlated-noise (CN) template [19]. This CN

contribution represents residual scale-dependent noise correlations in the component-separated y -map spectra, which become important on small angular scales. Hence, the model for the observed y -map power spectrum is

$$C_{\ell}^{yy} \simeq C_{\ell}^{\text{tSZ,1h}} + A_{\text{CIB}} C_{\ell}^{\text{CIB}} + A_{\text{IR}} C_{\ell}^{\text{IR}} + A_{\text{RS}} C_{\ell}^{\text{RS}} + A_{\text{CN}} C_{\ell}^{\text{CN}}, \quad (22)$$

where C_{ℓ}^{CIB} , C_{ℓ}^{IR} , C_{ℓ}^{RS} and C_{ℓ}^{CN} are templates of the foreground residuals and correlated noise terms, while the coefficients A_{CIB} , A_{IR} , A_{RS} , and A_{CN} are nuisance amplitudes. The templates are normalized to the fiducial residual levels estimated in the *Planck* y -map power-spectrum analysis. In practice, however, the residual foreground levels are uncertain and the different components are partially degenerate with each other and with the tSZ signal.

Traditional parameter inference methods for the tSZ power spectrum typically assume a Gaussian likelihood, given by $-2 \ln \mathcal{L} = \chi^2 + \ln |\mathbf{M}| + \text{const.}$, with

$$\chi^2 = (\hat{\mathbf{C}} - \mathbf{C})^T \mathbf{M}^{-1} (\hat{\mathbf{C}} - \mathbf{C}), \quad (23)$$

where $\hat{\mathbf{C}}$ and $\mathbf{C}$ are vectors containing the observed and theoretical tSZ power spectra $D_{\ell} = \ell(\ell+1)C_{\ell}/(2\pi)$ evaluated at all effective multipoles, and $\mathbf{M}$ is the fixed covariance matrix calculated using Eq. (20) evaluated at fixed cosmology. The posterior distribution of the model parameters is then sampled within a Markov chain Monte Carlo framework, typically using the Metropolis–Hastings (MH) algorithm.

However, as we have seen in Fig. 3, this Gaussian likelihood assumption fails to capture the skewness of the tSZ power spectrum distribution at low multipoles. A traditional Gaussian likelihood analysis may not provide an accurate description of the distribution of the data and may potentially lead to misestimated parameter constraints. This motivates considering alternative methods beyond the Gaussian likelihood analysis.

To evaluate the theoretical tSZ power spectrum and the covariance matrix at each point in parameter space, we use our Python package `tszsbi`.¹ It computes the 1-halo tSZ power spectrum and the trispectrum contribution to the covariance using `CosmoPower` emulators [36,37] that are used to build `class_sz` [38,39]. Written in JAX [40], `tszsbi` supports automatic differentiation and Just-In-Time compilation for fast likelihood evaluation.

V. FORWARD MODELING

Simulation-based inference allows us to infer model parameters using forward-modeled simulations without sampling from an explicit likelihood. To ensure the

parameter constraints are unbiased, a robust forward-modeling pipeline must be constructed to produce realistic mock observations across the parameter space of interest. In this section, we will first introduce a Gaussian likelihood-based simulator, which serves as the first validation of our SBI pipeline. We then move to a more realistic halo-based forward model, which captures the non-Gaussian statistics of the tSZ power spectrum.

A. Forward modeling for the Gaussian likelihood

As a first demonstration, we formulate the Gaussian likelihood analysis described in Sec. IV in an SBI-equivalent setting. We begin by constructing a forward simulator to simulate C_{ℓ}^{yy} for given input of cosmological and foreground parameters. We assume the probability distribution of the tSZ power spectrum follows a multivariate Gaussian distribution. The observational uncertainties can be modeled by adding Gaussian noise using Cholesky decomposition of the fixed covariance matrix, i.e., $\mathbf{M} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower-triangular matrix. The total simulated power spectrum is given by $C_{\ell}^{yy,\text{sim}} = C_{\ell}^{yy,\text{signal}} + N_{\ell}$, where $C_{\ell}^{yy,\text{signal}}$ is given by Eq. (22), and $\mathbf{N} = \mathbf{L}\mathbf{v}$, where $\mathbf{v}$ is a vector with elements drawn from a standard normal distribution. This method enables us to generate parameter-simulation pairs consistent with the Gaussian likelihood analysis, to which SBI can then be applied.

B. Forward modeling using a halo-based approach

To capture realistically the non-Gaussianity of the tSZ power spectrum due to massive and local clusters, we adopt a halo-based forward-modeling pipeline. In this pipeline, we first generate mock cluster catalogs for a given set of cosmological parameters, and then paint the clusters onto sky maps at random positions using their corresponding pressure profiles. Mock halo catalogs are generated by Poisson sampling of halo masses and redshifts from the HMF of Tinker *et al.* [30] using the `cosmocnc`² package [41]. The mass and redshift ranges of the halos are chosen to be $M \in [10^{14}, 10^{16}]M_{\odot}$, and $z \in [0.005, 3]$. Note that we adopt a higher minimum mass threshold than that used in the likelihood analysis of the real *Planck* data in Bolliet *et al.* [20], as the large-scale ($\ell < 10^3$) tSZ power is dominated by massive clusters, and the power spectrum converges rapidly for this M_{min} , as indicated in Sec. II C. This also significantly reduces the computational cost in the subsequent map-generation step, since the computational time of generating sky maps from halo catalogs scales with the number of clusters as $O(N_{\text{clusters}})$. Given the steep low-mass slope of the HMF, excluding low-mass halos substantially reduces the number of clusters in the catalogs while causing only a slight loss of tSZ power at high

¹<https://github.com/licongxu/tszsbi>

²<https://github.com/inigozubeldia/cosmocnc>

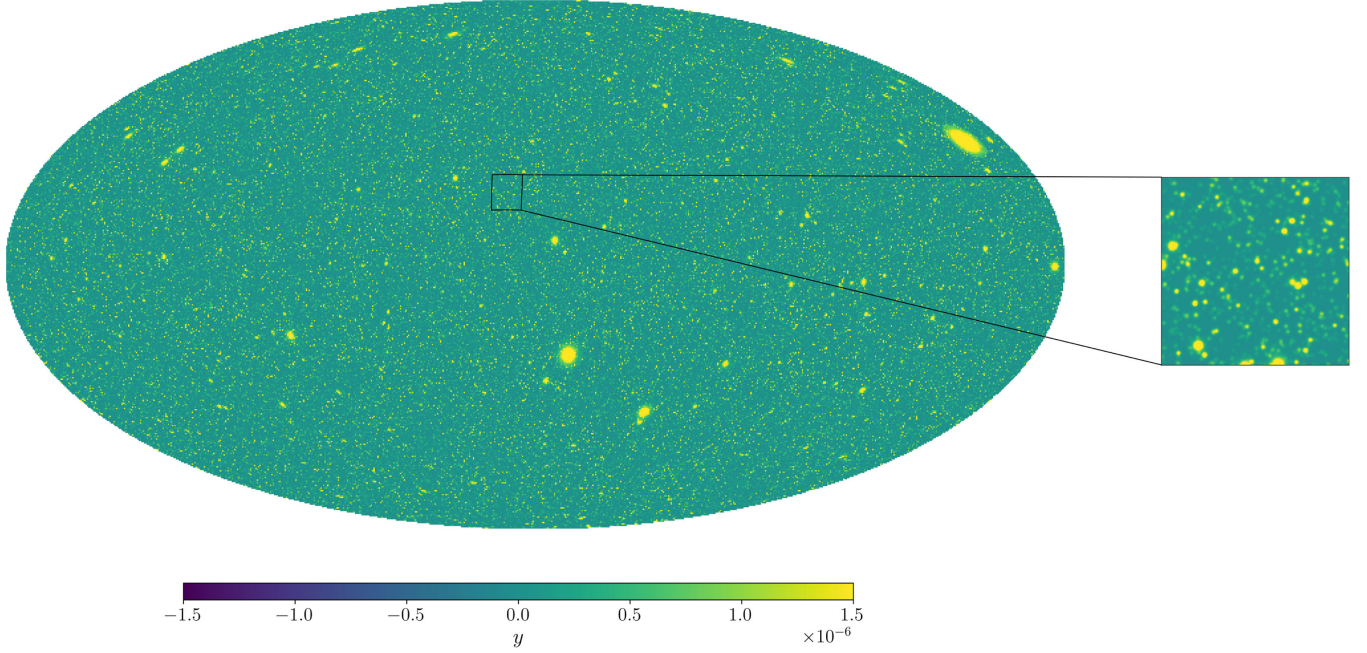


FIG. 4. Example of a noiseless full-sky Compton- y map and a $10^\circ \times 10^\circ$ patch extracted from it. The maps are generated using a modified version of XGPaint with $\theta_{\text{FWHM}} = 10$ arcmin and $N_{\text{side}} = 1024$.

multipoles, where the signal is already dominated by foreground residuals and instrumental noise.

The HMF is evaluated on a grid of $n_M = 5000$ mass points and $n_z = 5000$ redshift points. The total number of clusters in each catalog is then drawn from a Poisson distribution with mean given by the integral of the HMF over the chosen mass and redshift ranges. The (M, z) pairs of clusters in each catalog are subsequently sampled from the continuous distribution given by the HMF. Halo positions are sampled uniformly across the sky. We validate our synthetic catalogs by comparing the mean number counts across mass and redshift using 100 synthetic catalogs at fixed cosmology and the theoretical prediction of the input HMF from `cosmocnc`, as indicated in the Appendix B 1. No significant biases are found.

The full-sky maps are generated with a modified version of the XGPaint package.³ Each halo in the catalog is assigned the A10 pressure profile and convolved with a Gaussian beam with $\theta_{\text{FWHM}} = 10$ arcmin. The pressure profile is truncated at an angular radius equal to the larger of $4\theta_{500}$ or $2\theta_{\text{FWHM}}$, where $\theta_{500} = r_{500c}/d_A$. The choice of truncation radius is motivated by the fact that for clusters larger than the beam ($\theta_{500} \gtrsim \theta_{\text{FWHM}}$), one observes the cluster profile, whereas for clusters smaller than the beam ($\theta_{500} \lesssim \theta_{\text{FWHM}}$), one effectively observes the beam. Note that here we paint the beam-convolved cluster profiles directly onto the map, rather than painting the raw profiles and subsequently beam-convolving the entire map. In principle, the two methods are equivalent in the continuum

limit with no pixelization and in the absence of the truncation radius. In practice, however, each cluster must be painted over a finite angular support for computational efficiency. For small clusters with angular sizes comparable to or smaller than the map pixels, painting the beam-convolved profile therefore incorporates beam smoothing while ensuring compact clusters are sufficiently sampled on the pixel grid.

The maps are constructed on `Healpix.jl` [42] grids with resolution $N_{\text{side}} = 1024$. An example of the resulting noiseless full-sky Compton- y map and a $10^\circ \times 10^\circ$ patch extracted from it are shown in Fig. 4. Foreground residuals are added separately as Gaussian random fields generated from their residual beam-convolved power spectra. The angular power spectra are computed from these maps using the *Planck* 2015 binning scheme, with bandpowers defined at the 18 effective multipoles ℓ_{eff} of Ref. [19]. Within each bin, we apply uniform weights to D_ℓ when averaging over multipoles.

To validate the sky maps generated, we produce 1000 realizations at the fiducial cosmology and hydrostatic bias parameter, and compute the mean and covariance of the resulting tSZ power spectra. Figure 5 shows the mean and standard deviation of the tSZ power spectrum computed from 1000 simulated maps, as well as the covariance of C_ℓ between multipoles. The simulated mean spectrum is consistent with the theoretical prediction to within a few percent over the full multipole range, with subpercent-level agreement achieved at high multipoles, where the variance is small and most of the constraining power is concentrated. The covariance matrix obtained from simulations agrees

³<https://github.com/WebSky-CITA/XGPaint.jl>

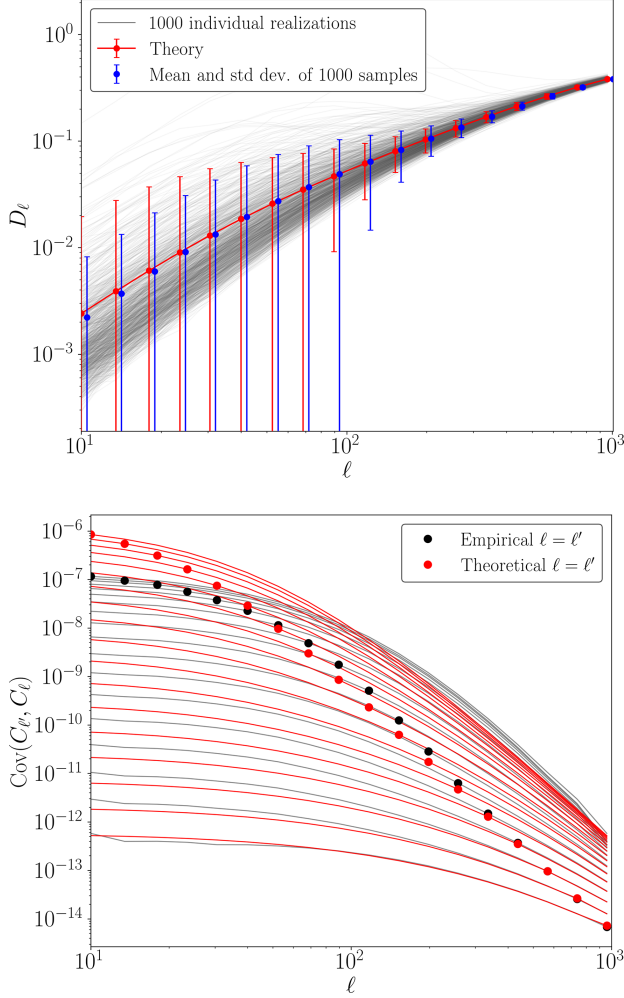


FIG. 5. Upper panel: Mean and standard deviation of the tSZ power spectrum computed from 1000 simulated maps (blue) compared with the corresponding theoretical predictions (red). The power spectrum of each individual realization is shown in gray. The empirical standard deviation agrees well with the theoretical error bars at high ℓ , while larger fluctuations are seen at low ℓ , where the variance is less well estimated from a finite number of simulations. Lower panel: Covariance matrix estimated from 1000 simulations (gray) and the theoretical prediction (red). Each line represents a different multipole ℓ' in the range 10–1000 (top to bottom). Filled circles indicate the diagonal elements of the covariance matrix. The mean simulated tSZ power spectrum agrees with the theoretical prediction overall, while showing larger scatter at low multipoles, as expected from the stronger sample variance in this regime.

well with the theoretical prediction at high multipoles. The agreement is poorer at low multipoles. This is because the variance of the error in the empirical covariance due to the finite number of simulations depends on the fourth connected moment of the sampling distribution of the C_ℓ . At low multipoles, the fourth connected moment is much larger than the square of the variance due to the heavy tails (see Fig. 3), giving large scatter in our empirical estimates.

VI. SBI IMPLEMENTATION

We use the `sbi` package [43], which implements SBI algorithms using neural networks built in PyTorch. In this paper, we consider both NPE [26] and NLE [27] to infer the posterior distribution. NPE models the posterior directly from parameter–data pairs, while NLE learns the likelihood, requiring an additional sampling step to obtain the posterior. In both cases, the neural networks are trained on a set of simulated data and corresponding cosmological and nuisance parameters, allowing efficient inference on the observed data.

We generate 9000 simulations for the likelihood-based SBI approach described in Sec. V A, and 18,000 simulations for the halo-based SBI method using our pipeline described in Sec. V B. Each SBI model is trained using an 80:20 train–validation split and optimized with the AdamW optimizer [44]. To stabilize training and prevent overfitting, we apply gradient clipping with a maximum norm of 5.0 and implement early stopping, terminating training if no improvement in validation loss is observed after 50 epochs. We adopt a Masked Autoregressive Flow (MAF) architecture to model the posterior or likelihood distribution [45]. We also explored other density-estimator architectures (e.g., Neural Spline Flows [46]) but found that they underperformed compared to the MAF. Instead, performance was dominated by hyperparameter tuning. In particular, the number of transformation layers and the size of the hidden layers were treated as hyperparameters and optimized to minimize the validation loss using wandb [47]. The adopted hyperparameter values are summarized in Table I.

We also find that training a single network repeatedly can lead to noticeable fluctuations in the inferred posterior, even when the same architecture and hyperparameters are used [48]. This can be due to the stochastic elements in training, such as random initialization and optimization paths. To mitigate instabilities arising from stochasticity in neural network training, we employ an ensemble of 10 independently trained networks and form a uniformly weighted mixture of their posteriors or likelihoods. By averaging the posteriors predicted by the ensemble, we obtain a more stable and robust estimate of the inferred posterior distribution.

VII. RESULTS

In this section, we present the results of our parameter constraints using both the likelihood-based and halo-based SBI approaches. We first compute the parameter constraints by sampling from the Gaussian likelihood using Cobaya [49,50] with the MH algorithm, and compare with those obtained from its SBI-equivalent settings to validate the SBI application in this task. We then compare the results obtained by halo-based SBI with those from the Gaussian likelihood to assess how the inferred parameter constraints change when the Gaussian-likelihood assumption is replaced by a halo-based forward model.

TABLE I. Training hyperparameters for the MAF architecture used in the likelihood-based and halo-based SBI approaches. The table lists the neural network and optimization settings for both NPE and NLE approaches.

Hyperparameter	NPE (likelihood-based)	NLE (likelihood-based)	NPE (halo-based)	NLE (halo-based)
Hidden features	80	20	150	30
Number of transforms	15	6	10	18
Number of components	5	5	5	5
Number of bins	3	3	3	3
Dropout probability	0.05	0	0.05	0
Activation function	ReLU	ReLU	tanh	tanh
Batch size	64	128	256	256
Initial learning rate	2.382×10^{-4}	5.593×10^{-3}	4.871×10^{-4}	1.758×10^{-3}
Learning rate decay factor (γ)	0.9528	0.8304	0.8551	0.7341
Scheduler step size	43 epochs	22 epochs	63 epochs	36 epochs

A. Data

In this work, we use two distinct data setups. For the likelihood-based analysis, we use the real power spectrum data measured by *Planck*, with multipole bins ranging from $10 \leq \ell \leq 959.5$ following Bolliet *et al.* [20]. The likelihood analysis is performed directly on these data using the Gaussian likelihood described in Sec. IV, while the corresponding SBI comparison is trained on samples drawn from the same Gaussian likelihood with the noise drawn from the Cholesky decomposition of the covariance matrix (Sec. VA), so that it provides an SBI equivalent of the likelihood-based setup. By contrast, for the halo-based analysis, we instead define a mock observed power spectrum from a single realization of the halo-based forward model at the fiducial parameters as described in Sec. VB, and then compare halo-based SBI constraints against those obtained by applying the Gaussian-likelihood analysis to this same mock dataset. We neglect instrumental noise ($A_{\text{CN}} = 0$) in this setup. These steps allow us first to validate SBI against the standard likelihood-based approach on real data, and then assess how the inferred constraints change when the Gaussian-likelihood assumption is replaced by a halo-based forward model in a controlled mock-data case. In both setups, we vary the

TABLE II. Priors imposed on cosmological and foreground parameters in likelihood analysis. We adopt uniform priors on all parameters, where $\mathcal{U}(a, b)$ denotes a uniform distribution between a and b .

Parameter	Prior distribution
$\Omega_b h^2$	$\mathcal{U}(0.02, 0.025)$
$\Omega_c h^2$	$\mathcal{U}(0.11, 0.13)$
H_0 (km s $^{-1}$ Mpc $^{-1}$)	$\mathcal{U}(55, 90)$
$\ln(10^{10} A_s)$	$\mathcal{U}(2.5, 3.5)$
n_s	$\mathcal{U}(0.94, 1.0)$
B	$\mathcal{U}(1.0, 2.0)$
A_{CIB}	$\mathcal{U}(0, 5)$
A_{IR}	$\mathcal{U}(0, 5)$
A_{RS}	$\mathcal{U}(0, 5)$

five cosmological parameters, the hydrostatic bias parameter B , and the three residual foreground amplitudes with the priors shown in Table II.

B. Constraints from likelihood-based analysis

Figure 6 compares the posterior constraints obtained from the Gaussian-likelihood analysis using the MH

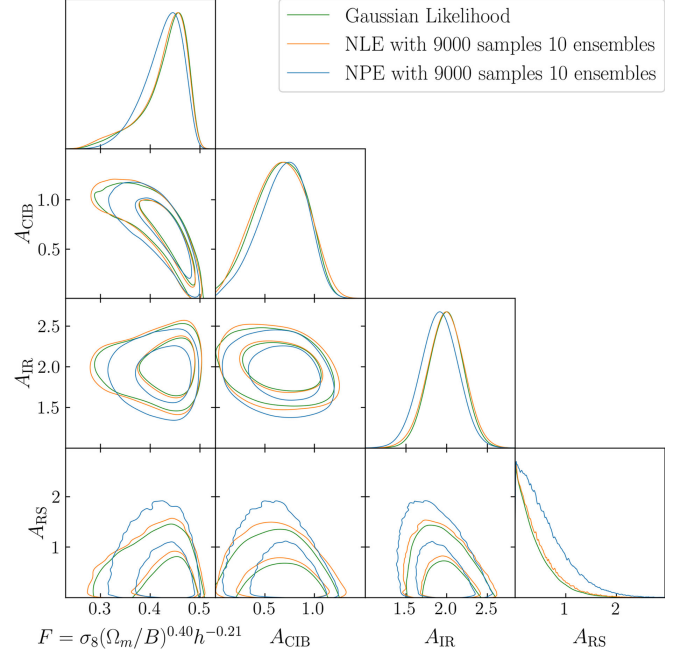


FIG. 6. Posterior constraints from the Gaussian likelihood analysis and the SBI analyses applied to the real *Planck* data measured by Bolliet *et al.* [20]. The green contours show the Gaussian likelihood result, while the orange and blue contours show the NLE and NPE results, respectively. Both SBI analyses are trained using synthetic data generated with the same Gaussian power-spectrum forward model assumed in the likelihood analysis. Comparing posterior means of each parameter, the NLE results agree with the likelihood results to within 0.05σ , while the NPE results agree to within 0.4σ , with the largest mean shift occurring for A_{IR} . Contours indicate the 68% and 95% credible regions.

TABLE III. Comparison of parameter constraints from the Gaussian likelihood analysis and NPE and NLE SBI-equivalent analyses using real *Planck* data measured by Bolliet *et al.* [20]. The posterior distributions are shown in Fig. 6. The quoted parameter constraints correspond to 68% credible bounds.

Parameter	MH	NLE	NPE
F	$0.430^{+0.054}_{-0.018}$	$0.428^{+0.055}_{-0.018}$	$0.425^{+0.049}_{-0.025}$
A_{CIB}	$0.66^{+0.30}_{-0.23}$	$0.65^{+0.30}_{-0.25}$	$0.68^{+0.28}_{-0.20}$
A_{IR}	1.99 ± 0.21	1.99 ± 0.22	1.91 ± 0.23
A_{RS}	$0.39^{+0.073}_{-0.39}$	$0.422^{+0.090}_{-0.42}$	$0.58^{+0.13}_{-0.57}$

sampler with those inferred using SBI methods under the same Gaussian likelihood assumption. For the MH reference analysis, convergence is assessed using the Gelman–Rubin criterion $\hat{R} - 1 < 0.01$ [51]. The corresponding parameter constraints are summarized in Table III. We find excellent agreement between the NLE posterior and the MH reference posterior, with the differences in the mean inferred values being less than 0.05σ . The NPE posteriors show poorer agreement with the likelihood analysis, with the largest deviation being about 0.4σ for A_{IR} . This likely reflects the intrinsic difficulty for NPE to learn the highly non-Gaussian posterior with a MAF neural network architecture. This difficulty may arise from degeneracies between the tSZ power spectrum signal and the foreground templates, which make the posterior highly non-Gaussian. The positivity prior on A_{RS} may also contribute by truncating part of the posterior support, whereas NLE is unaffected by such boundary effects, although in principle the MAF model should be able to learn such prior-induced boundaries. We note that the contribution from radio sources is poorly constrained compared to the cosmological constraint F and other foreground residual amplitudes. This is primarily due to degeneracies between the radio-source template and the other foreground components and the tSZ signal, as also noted in Planck Collaboration [19]. Nevertheless, we find excellent agreement between the NLE and MH posteriors, despite the template degeneracies leading to numerical difficulties. This demonstrates the ability of SBI to produce reliable posterior estimates for this density estimation task.

C. Constraints from halo-based SBI

We next turn to the halo-based forward-modeling setup, with a simple scenario in which instrumental noise is neglected. The data vector consists of the tSZ power spectra computed from maps generated from the simulated halo catalogs, to which we add residual foregrounds (CIB, IR, and RS) drawn as Gaussian random fields. For benchmarking purposes, we take as our data a single realization of the tSZ map drawn from the 1000 simulations described in Sec. VB, combined with one realization of the foreground residuals generated with parameters $A_{\text{CIB}} = 0.66$,

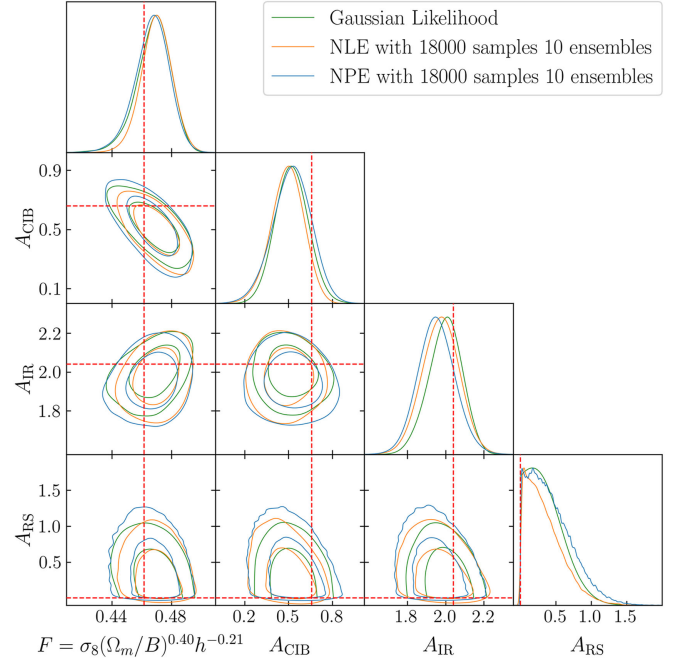


FIG. 7. Parameter constraints from a simulated halo-based realization of the tSZ power spectrum. Green contours are for an approximate Gaussian-likelihood analysis while SBI results are shown for NLE (orange) and NPE (blue). Red dashed lines indicate the true parameter values used to generate the benchmark simulation. Contours indicate the 68% and 95% credible regions.

$A_{\text{IR}} = 2.04$, and $A_{\text{RS}} = 0.0004$, which are chosen to be consistent with the foreground-amplitude constraints obtained from the *Planck* analysis in Sec. VII B.

Figure 7 shows the posterior contours obtained from both SBI approaches and the (approximate) Gaussian likelihood. The corresponding parameter constraints are summarized in Table IV. All three methods recover constraints that are consistent with the input parameter values. Remarkably, both the NLE and NPE approaches yield the posterior distributions for the cosmological parameter F in excellent agreement with the likelihood-based result, despite the fact that the likelihood is highly non-Gaussian at low ℓ . However, there are some differences in the foreground residual amplitude constraints from SBI

TABLE IV. Comparison of parameter constraints from a Gaussian likelihood (MH) and SBI with NLE and NPE applied to a simulated halo-based realization of the tSZ power spectrum. The posterior distributions are shown in Fig. 7. All quoted parameter constraints correspond to 68% credible bounds.

Parameter	MH	NLE	NPE
F	$0.468^{+0.013}_{-0.0097}$	0.470 ± 0.010	$0.468^{+0.012}_{-0.0098}$
A_{CIB}	0.52 ± 0.11	$0.49^{+0.12}_{-0.11}$	$0.51^{+0.14}_{-0.13}$
A_{IR}	$1.998^{+0.094}_{-0.081}$	1.974 ± 0.096	1.955 ± 0.095
A_{RS}	$0.37^{+0.094}_{-0.37}$	$0.35^{+0.10}_{-0.33}$	$0.358^{+0.094}_{-0.35}$

and the likelihood-based method. We note that the SBI-based methods yield slightly larger uncertainties in all three amplitudes compared to the likelihood analysis. The constraint on A_{CIB} from NLE is 0.3σ lower than the likelihood-based result, while the A_{IR} value obtained from NPE is 0.5σ lower than the likelihood-based constraint. For A_{RS} , both SBI methods yield the inferred means below the mean inferred from the likelihood-based analysis.

Based on the benchmark on the real *Planck* data in Sec. VII B, where NLE showed closer agreement with the MH reference posterior, we regard the NLE posterior as the more robust SBI estimate in this test. This is further supported by the validation test in Sec. VIII, which shows that NLE can accurately learn the non-Gaussian likelihood structure across the multipole range (see Figs. 11 and 12 there). The discrepancy between NLE and NPE posteriors may arise from NPE-specific limitations, including incomplete convergence (see Fig. 9 later) the limited network architecture capacity, and the positivity prior on A_{RS} as discussed in the likelihood benchmark test in Sec. VII B.

Overall, despite the true likelihood being non-Gaussian, the SBI-based constraints on the cosmology-derived parameter F are in excellent agreement with those obtained from the Gaussian likelihood analysis. The discrepancies in the foreground residual constraints may arise from several factors. For example,

- (i) the true likelihood certainly differs from the Gaussian approximation due to the intrinsic non-Gaussianity of the signal;
- (ii) the SBI-related artifacts, including incomplete convergence due to a finite number of simulations and stochasticity in neural network training; and
- (iii) mismatches between the simulations and the theoretical model, since the SBI posterior is learned entirely from XGPaint simulations, while the likelihood analysis is based on theory and applied to one realization generated by the same simulation pipeline.

The discrepancies are most likely driven by the first factor, since the power spectrum remains slightly skewed even at high multipoles. SBI-related artifacts are less likely given our benchmark experiment in Sec. VII B, which shows good agreement between SBI and the likelihood-based analysis for the same kind of data under the Gaussian-likelihood assumption. Nevertheless, we carry out further validation tests to assess possible SBI artifacts and simulation biases in Sec. VIII.

D. Reconstructed tSZ power spectrum

Figure 8 shows the reconstructed power spectra of the tSZ signal and foreground residuals from the halo-based setup in Sec. VII C for the NLE method. The black dots denote the total power spectrum of the mock observation, while the gray violin plots show the sampling distribution of the 1000 simulated bandpowers in each multipole bin.

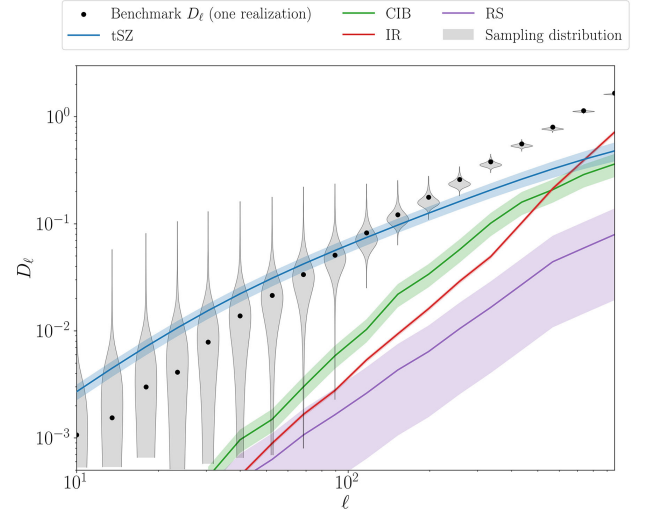


FIG. 8. Reconstructed tSZ power spectrum and foreground residuals for a single simulated realization using the NLE method. The black dots represent the total binned power spectrum computed from one realization of simulated maps with tSZ, CIB, IR and RS included. The gray violin plots show the sampling distribution of the 1000 simulated bandpowers in each multipole bin, illustrating the positively skewed distribution of the power spectrum, especially at low multipoles. The blue curve and shaded band show the mean and 1σ uncertainty of the reconstructed tSZ spectrum obtained from 1000 posterior samples. Foreground residuals are shown as template-shaped curves with correlated amplitude uncertainties translated into smooth shaded bands. The uncertainties are represented as smooth shaded bands, because all amplitude uncertainties (for both tSZ and foregrounds) create fully correlated fluctuations across multipoles.

This ensemble of simulated bandpowers is generated at fixed cosmological and residual foreground parameters, matching the mock observation. To obtain the reconstructed tSZ signal, we draw 1000 samples from the posterior distribution of the parameters and compute, at each multipole, the theoretical 1-halo tSZ power spectrum for each sample. Since we use fixed templates for foreground residuals, the mean and standard deviation of each foreground amplitude translate directly into the mean and uncertainty band of the corresponding foreground residual power spectrum. Because all amplitude uncertainties (for both tSZ and foregrounds) are fully correlated across multipoles, we represent the reconstructed uncertainties using smooth shaded bands rather than independent error bars. The reconstructed tSZ spectrum is detected with a mean signal-to-noise ratio of 5.6. This is consistent with the posterior constraint on F in Table IV with a fractional uncertainty of around 2.4%. Since F controls the overall amplitude of the tSZ power spectrum and the mean tSZ power scales approximately as F^8 , this corresponds to a fractional uncertainty of about $8 \times 2.4\% = 19\%$ in the reconstructed tSZ power spectrum, or a signal-to-noise ratio of 5.2, broadly consistent with the measured value of 5.6.

VIII. VALIDATION OF THE SBI METHODS

A. Convergence of SBI with number of simulations

To assess whether the small differences in the inferred constraints for the halo-based setup (Fig. 7) are caused by genuine non-Gaussianity in the likelihood or from training and simulation artifacts, we perform a set of ablation studies. We first consider the convergence of the SBI methods as a function of the number of simulations. To do so, we examine the behavior of the loss function with respect to the training sample size. The loss function is the negative log probability used to train the neural density estimator on simulated parameter–data pairs $\{(\theta_i, \mathbf{t}_i)\}$, where $\mathbf{t}$ denotes the summary statistics of the data. For NPE, this is the negative log posterior, $-\log p_\phi(\theta_i|\mathbf{t}_i)$, while for NLE it is the negative log likelihood, $-\log p_\phi(\mathbf{t}_i|\theta_i)$, where ϕ denotes the trainable parameters of the neural density estimator. The validation loss is the same quantity evaluated on simulations not used for training, and therefore provides a diagnostic of generalization.

Figure 9 shows the best validation loss as a function of the number of simulations used to train the NLE and NPE neural networks, respectively, plotted with a logarithmic x -axis. All training runs are performed using the same hyperparameters listed in Table I for each training set. For each training size between 500 and 20,000 simulations, we also tested different neural-network hyperparameters and found that the results are only weakly affected by this choice. For NLE, the validation loss decreases steeply at small training sizes and plateaus beyond approximately 8000 simulations, indicating that the model has converged. By contrast, for NPE the loss decreases approximately linearly on the log–linear plot across the full range of training sizes explored. This power-law-like trend may suggest that NPE would require substantially more simulations to reach convergence. This is possibly due to the difficulty of learning the highly non-Gaussian posterior given our uniform priors under limited data, or to limitations of the MAF architecture when learning such complex posterior distributions.

We further examine the convergence of the parameter means and standard deviations obtained from the posterior as the training sample size increases. Figure 10 shows the convergence of the posterior means and standard deviations as a function of the number of simulations used for training. Since this process is computationally heavy (especially for NLE), we draw the posterior from an ensemble of only three independently trained neural networks for each training size (we used 10 for our earlier results). For all four parameters, the posterior means and standard deviations converge after about $N \simeq 10,000$ training samples. However, some residual scatter remains for individual parameters, most notably for A_{CIB} , where fluctuations can reach up to roughly 0.5σ . This may arise from several factors, such as the stochasticity in training and the finite size of the training dataset (even though the NLE validation

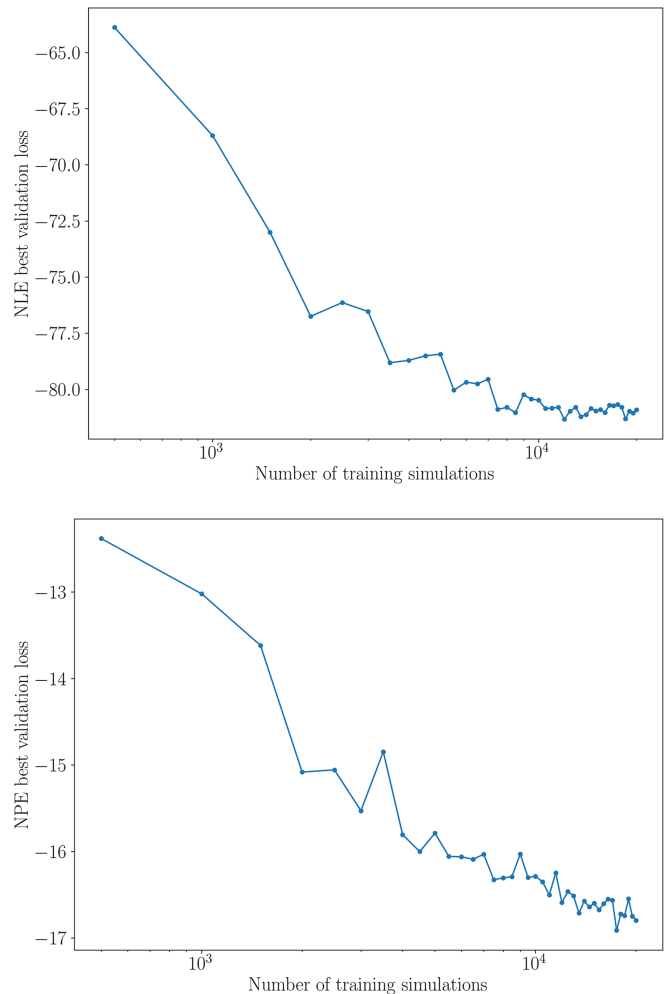


FIG. 9. Best validation loss as a function of the number of training simulations. Upper panel: NLE, converging after around 7500 simulations. Lower panel: NPE, converging more slowly and approaching stability after around 17,500 simulations.

loss has converged, whereas the NPE loss has not; see Fig. 9). Apart from that, fluctuations in the Poisson tail of the simulated data used in training may also contribute to the observed scatter.

B. Learned likelihood (NLE)

One benefit of NLE is that we can validate the learned likelihood for both the Gaussian likelihood-based and halo-based SBI frameworks. We compare the learned likelihood from the SBI neural network with the actual sampling distribution of the data obtained from the simulations.

For the SBI setup based on the Gaussian likelihood, we draw 10^5 samples from the learned likelihood conditioned on the fiducial cosmological and residual foreground parameter values. We compare these samples with the 10^3 independent realizations of D_ℓ^{yy} reserved as the held-out test set, generated at the same parameter values using different noise realizations drawn from the theoretical

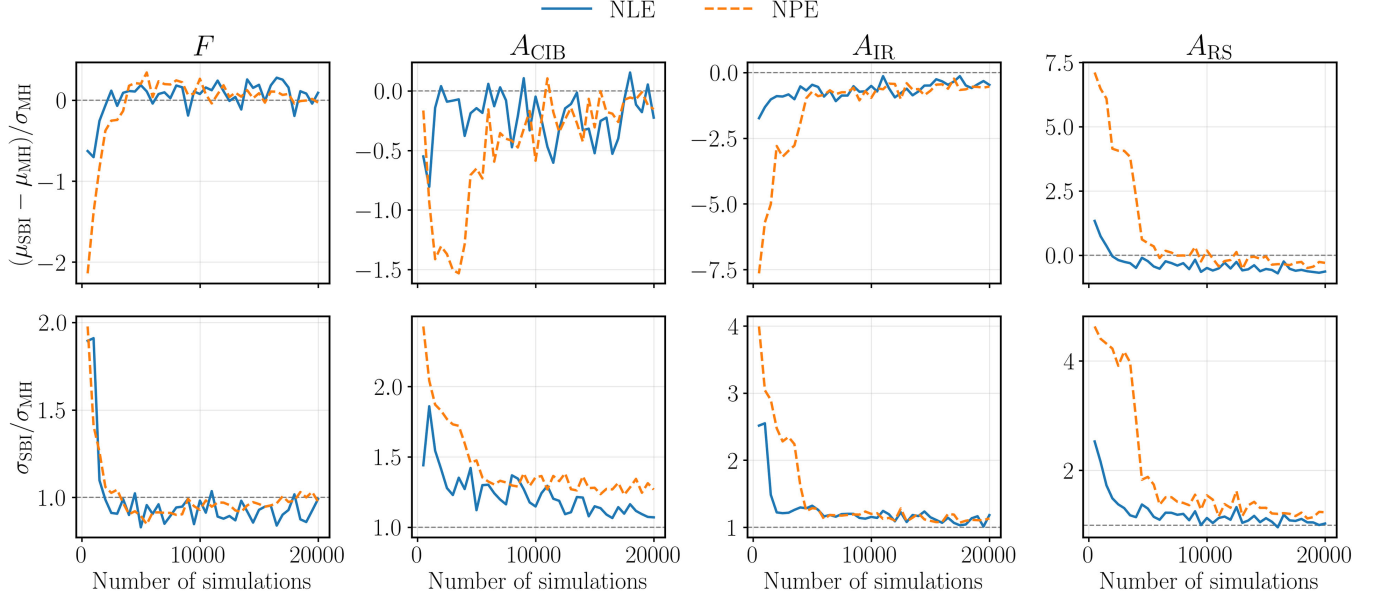


FIG. 10. Convergence of the posterior means (top panels) and standard deviations (bottom panels) as a function of the number of training simulations for all four parameters. Results for NLE are shown in solid blue and NPE in dashed orange. Each point represents the result from an ensemble of three independently trained neural networks. In all cases, the SBI results are referenced to the results from a Gaussian-likelihood analysis.

covariance matrix. These test realizations are not used during training or validation, and provide an independent estimate of the true sampling distribution of the data vector. Figure 11 shows the resulting histograms, together with

the Gaussian prediction derived from the theoretical covariance matrix. The agreement is excellent, indicating the reliability of the NLE approach in capturing the underlying distribution.

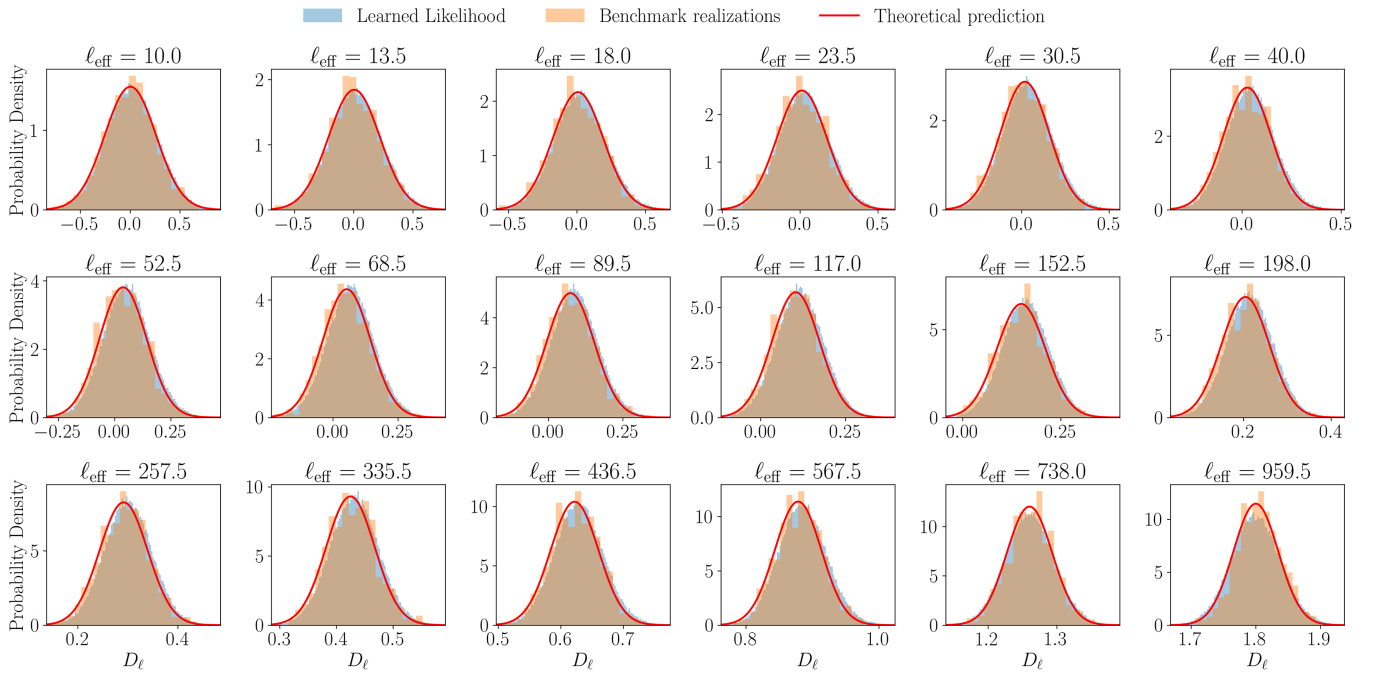


FIG. 11. Histograms showing the distributions of the tSZ power spectrum drawn from the learned likelihood (blue) by assuming a Gaussian distribution at all multipoles, and the distributions of the tSZ power spectrum from 1000 different draws from the covariance matrix at the same parameter values (orange). The red curve indicates the Gaussian theoretical prediction from the mean and covariance matrix.

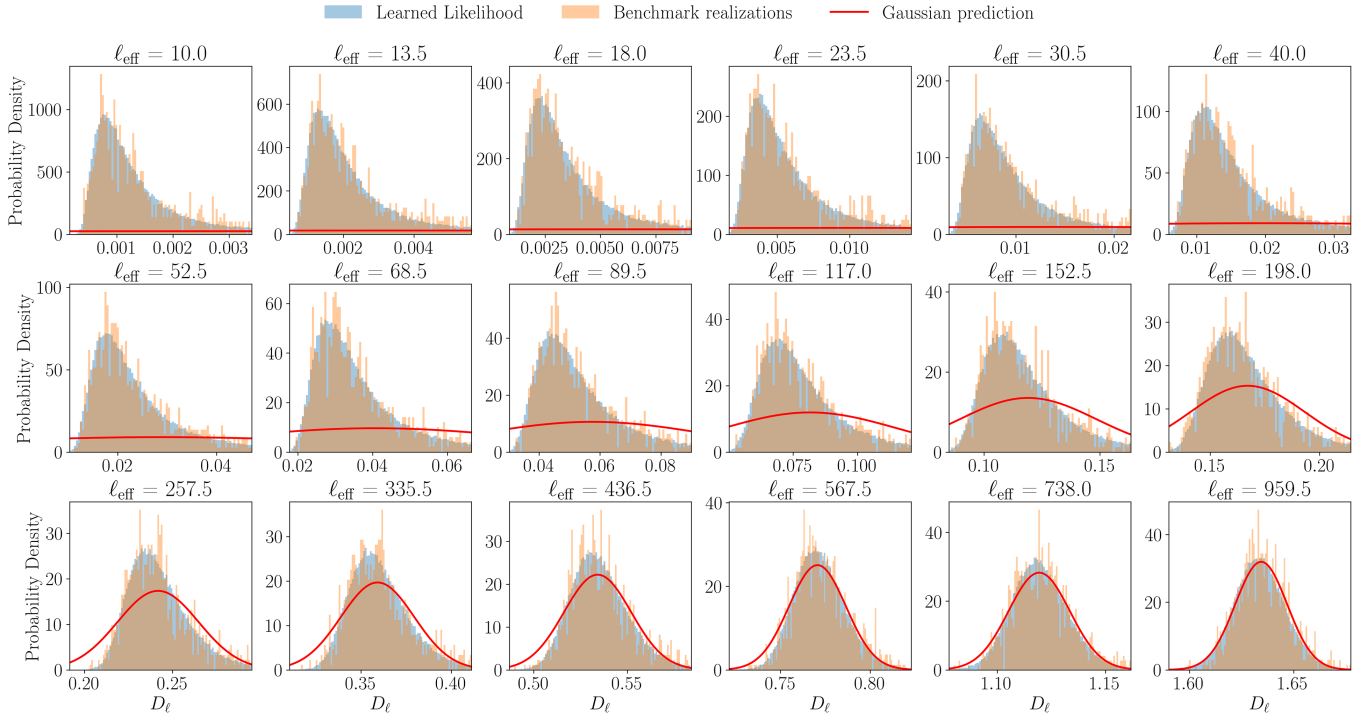


FIG. 12. Histograms showing the distributions of the tSZ power spectrum drawn from the learned likelihood (blue) and the sampling distributions of the tSZ power spectrum from 1000 forward-modeled realizations (orange). The red curve shows the Gaussian theoretical prediction derived from the mean and covariance matrix. At low multipoles, the Gaussian approximation fails to describe the strongly non-Gaussian distribution, whereas the learned likelihood successfully reproduces the skewness in the bandpower distribution.

Similarly, for the halo-based approach, we perform an analogous test of drawing 10^5 samples from the learned likelihood at fixed parameter values and comparing against 10^3 forward-modeled realizations of D_ℓ^{yy} . The same 1000 halo-based tSZ maps described in Sec. VB are used to generate these realizations. Figure 12 presents the resulting histograms and the Gaussian predictions. At low multipoles, the bandpower distributions are strongly positively skewed, and the Gaussian theory curves fail to capture this highly non-Gaussian behavior. In contrast, the samples drawn from the learned NLE likelihood closely match the held-out forward-modeled distributions, including their skewness.

This comparison is also useful for diagnosing possible sources of posterior bias. Since SBI is trained directly on the simulated data distribution, any systematic mismatch between the forward-modeled simulations and the theoretical prediction can be inherited by the learned likelihood and hence by the inferred posterior. At higher multipoles, where foreground residuals dominate and the distribution is closer to Gaussian, we find a small mean offset of approximately 0.1σ between the held-out forward realizations and the Gaussian prediction in bins with $\ell > 100$. This offset could therefore contribute marginally to the small shift in A_{IR} seen in the SBI parameter recovery tests in Sec. VII C. However, the size of the effect is too small to explain the full observed discrepancy: correcting for this

offset would change the inferred A_{IR} by only 0.007, compared to the observed shift of approximately 0.02.

C. Coverage test (NPE)

For the halo-based SBI case study using NPE, we perform a coverage test to assess whether the inferred posteriors are statistically consistent with the true parameters. Since NPE is amortized, the trained neural network can generate posteriors for arbitrary observations in a short time without retraining. In contrast, performing the same test for NLE is much more computationally expensive, as the posterior must be obtained by sampling from the learned likelihood for each individual observation. Moreover, the validation test for NLE described in Sec. VIII B is already robust. We construct 1000 independent test cases by sampling parameter sets from the prior, generating the corresponding y -maps and power spectra, and inferring posteriors for each using the trained network. For every test case, 10,000 posterior samples are drawn and compared to the corresponding true parameter values. We consider two complementary diagnostics. First, we compare the posterior samples with the true input parameters across the test set, which provides a direct visual check for biases in the inferred posterior. Second, we perform a credible-level coverage test, in which we check whether posterior credible regions of nominal probability p contain the true parameter values in approximately a fraction p of the test simulations.

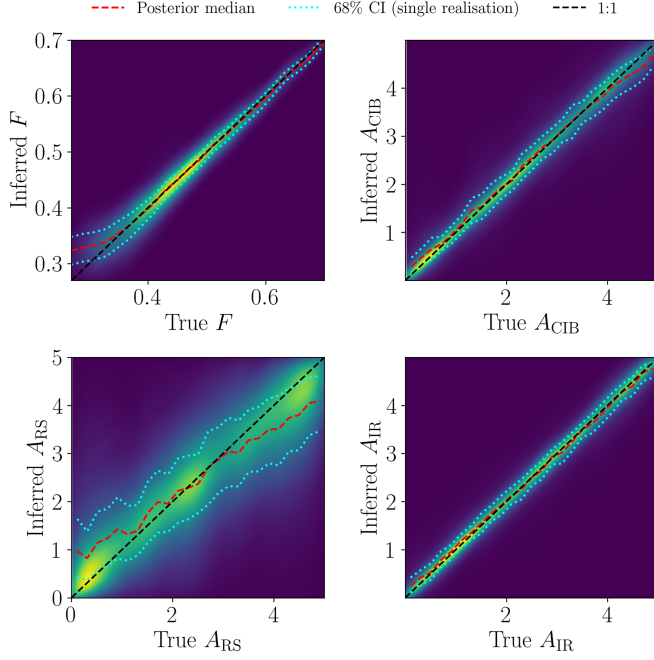


FIG. 13. Results of the coverage test comparing the input (true) parameter values (x -axis) to the inferred ones (y -axis), using a random subsample of 100,000 points from the posterior–truth dataset. The point density is smoothed using a Gaussian kernel density estimator and shown by the color scale, with the median relation indicated by a red dashed line and the 1:1 line shown in black. The cyan dotted curves show the central 68% credible interval for a single realization at each true parameter value, corresponding to the 16th and 84th posterior percentiles.

Figure 13 shows the two-dimensional density plots comparing the inferred posterior to the true parameter values for all four parameters. We visualize the result using a Gaussian kernel density estimator applied to a random subsample of 100,000 points from the full posterior–truth dataset. The indication that our NPE framework has good coverage is the fact that there is a high density of points around the 1:1 line for all four parameters, except at the edge of the prior. Among all four parameters, only A_{RS} remains poorly constrained, as expected given its weak contribution to the tSZ power spectrum.

Figure 14 shows the results of another coverage test. Here, for each of the posteriors corresponding to the 1000 independent prior samples and data realizations, we determine, using our trained NPE network, the highest posterior density region for each parameter enclosing probability $p/100$ (the nominal credibility⁴) and compute the fraction of these that include the true parameter value (the empirical coverage). This provides a frequentist calibration test of the

⁴We parametrize the nominal credibility by z_p through the two-sided Gaussian relation $p/100 = \frac{1}{\sqrt{2\pi}} \int_{-z_p}^{z_p} e^{-z^2/2} dz$, such that $z_p = 1, 2$, and 3 correspond to the conventional 68.27, 95.45, and 99.73 percent credible intervals, respectively.

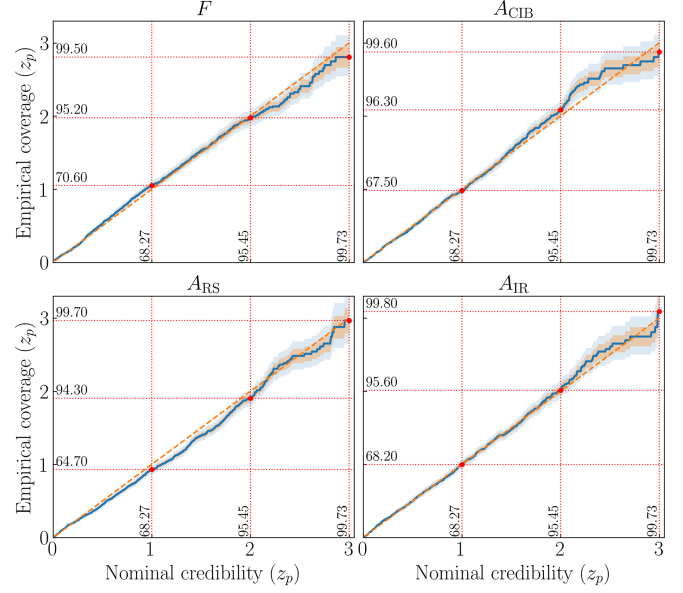


FIG. 14. Coverage test comparing the empirical coverage to the nominal credibility level z_p . The dark blue lines show the mean empirical coverage, the orange shaded regions indicate the 68% uncertainties, and the light-blue shaded regions correspond to the 95% uncertainties. The orange dashed lines denote the 1:1 relation. The deviation near the high nominal credibility is due to the tails of the learned posterior and the finite-sample fluctuations in the test realizations.

Bayesian credible regions: if the posterior is well calibrated, the empirical coverage should agree with the nominal credibility level, while empirical coverage below the nominal level indicates over-confident posteriors [52]. The empirical coverage and the nominal credibility levels generally match to good precision. The largest deviations occur at high nominal credibility, where the test is sensitive to the tails of the learned posterior and to finite-sample fluctuations in the test realizations.

In addition, we quantify the level of bias in the posteriors by computing the ratio between the difference in the inferred mean (μ_{SBI}) and the true value (μ_{true}) and the posterior standard deviation from SBI (σ_{SBI}) at those 1000 different parameter points. Table V summarizes this ratio for each parameter, averaged over 1000 different parameter points. For all parameters, $(\mu_{SBI} - \mu_{true})/\sigma_{SBI}$ is less than 0.06, with the largest deviation being 0.0534 for A_{IR} .

TABLE V. Difference between SBI-based posterior constraints and the true input parameters in units of σ_{SBI} . For all parameters, the bias is less than 0.06σ .

Parameter	$(\mu_{SBI} - \mu_{true})/\sigma_{SBI}$
F	0.0061 ± 0.0320
A_{CIB}	0.0457 ± 0.0309
A_{RS}	-0.0489 ± 0.0327
A_{IR}	0.0534 ± 0.0311

Taken together, our results indicate that the NPE posterior estimator is well calibrated across the prior range. Nevertheless, passing the ensemble-level coverage tests is not sufficient for precise cosmological parameter inference. The difference between the NLE and NPE results in Fig. 7 is comparable to the difference between these SBI estimates and the posterior obtained under the Gaussian-likelihood assumption. Such shifts are small compared with the quoted uncertainties and therefore may not significantly affect the coverage statistics, but they are large enough to be relevant for cross-checks against the Gaussian-likelihood results. The residual NPE discrepancies are also consistent with the incomplete convergence of the NPE validation loss. We therefore regard the NPE validation presented in this section as necessary but not sufficient. A more stringent diagnostic is provided by the NLE likelihood validation (Fig. 11 and 12), while interpreting the NPE coverage tests as approximate calibration rather than full numerical agreement with the likelihood benchmark.

IX. DISCUSSION AND OUTLOOK

In this paper, we revisited the theoretical modeling of the sampling distribution of the tSZ power spectrum, and implemented an SBI-based framework for constraining cosmological parameters from the tSZ power spectrum using halo-based simulations. We found that approximating with a Gaussian likelihood yields unbiased cosmological parameter estimates for a *Planck*-like survey even in the presence of the strongly non-Gaussian sampling distribution at low multipoles. However, the constraints on the amplitudes of the foreground residuals in the SBI analysis are slightly broader than those from the Gaussian likelihood analysis, and the inferred value of the amplitude of Poisson power from infrared sources, A_{IR} , is around 0.5σ lower relative to the likelihood-based result. We thoroughly validated the SBI implementation with a series of ablation studies in NLE and NPE to argue that this shift in A_{IR} is less likely due to inference systematics but rather reflects a genuine difference between the true and Gaussian likelihoods.

This is the first time that SBI has been applied to the analysis of the tSZ power spectrum. While a standard Gaussian likelihood approximation yields unbiased cosmological parameter estimates for a *Planck*-like survey considered here, this conclusion does not automatically hold for other surveys such as ACT and the South Pole Telescope (SPT). ACT and SPT have lower noise and a higher angular resolution, but they both observe a smaller fraction of the sky, which increases the sample variance and makes the likelihood of the tSZ power spectrum more non-Gaussian even at higher multipoles. Moreover, the tSZ power spectrum depends more strongly on the pressure profile and baryonic effects at smaller angular scales. Even if a Gaussian likelihood is sufficient to characterize the bandpower distributions, the inference can still be limited by model misspecification in the pressure profile rather

than likelihood non-Gaussianity. In this sense, SBI can offer an advantage in that it can be trained to incorporate uncertainties in the pressure profiles and small-scale astrophysical effects.

Our results also suggest that caution is required when interpreting the discrepancies between the observed tSZ power spectrum and simulation-based templates from hydrodynamical simulations, e.g., [53,54]. The bandpower distributions are positively skewed in the *Planck*-like regime studied here, so the likelihood is intrinsically asymmetrical as shown in Figs. 8 and 12. Hence, differences between measured bandpowers and a fiducial template do not map straightforwardly onto shifts in parameters. A direct visual comparison between measured bandpowers and a fiducial template can be misleading, and apparent deviations in bandpower space do not translate straightforwardly into parameter shifts.

It is worth noting that the non-Gaussianity of the tSZ power spectrum at low multipoles arises primarily from Poisson fluctuations due to a small number of local, massive clusters. Therefore, when those massive clusters are included, the tSZ power spectrum analysis is subject to large sample variance due to the trispectrum covariance. Masking these resolved clusters can therefore make the low-multipole regime more Gaussian and enhance the amount of cosmological information from the tSZ power spectrum; see, e.g., [55,56]. Meanwhile, the information in the tSZ power spectrum removed by masking is transferred to the catalog of resolved clusters. This motivates a joint analysis of the unresolved tSZ power spectrum and the number counts of the masked clusters [18]. While an analytical likelihood can provide an accurate approximation for each observable individually, the SBI framework naturally captures any covariance (especially if the large-scale clustering effect from the 2-halo term) and provides a convenient way to combine information from two different observables. We leave the further exploration of this to future work.

As a first demonstration, the present halo-based SBI implementation only considers full-sky coverage and neglects realistic noise. These effects can, in principle, be incorporated into the forward-modeling framework, although the accuracy of the resulting inference will depend on how well the noise properties can be realistically simulated. Extensions to incorporate realistic noise and survey masks will allow a reanalysis of current datasets such as *Planck* [19], and to produce forecasts for ongoing experiments such as Simons Observatory [57].

We finally emphasize that our forward-modeling approach is both universal and highly adaptable, providing a flexible framework for a wide range of applications beyond the specific tests presented here. While in this work we focused on the tSZ power spectrum as the summary statistic, the pipeline can equally be applied to other statistics or combined datasets. In particular, it naturally accommodates joint analyses with complementary probes such as cluster number counts [14], and higher-order

moments [10]. These extensions offer other possible directions for exploiting the full statistical power of current and forthcoming CMB surveys.

Figures 6 and 7 were produced using the `GetDist` package [58].

ACKNOWLEDGMENTS

We thank William Coulton and Zack Li for assistance with the use of the `XGPaint` package. L. X. acknowledges the support of China Scholarship Council Cambridge Scholarship (Grant No. 202408060222). Í. Z. and A. C. acknowledge support from the STFC (Grants No. ST/W000977/1, No. ST/X006387/1 and No. UKRI1164). J. A. is supported by a fellowship from the Kavli Foundation. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service [59], provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (Capital Grant No. EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council [60].

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX A: CONDITIONAL SCATTER OF THE TSZ POWER SPECTRUM AT FIXED HALO COUNTS

In Sec. III, we model the one-point distribution of the tSZ power spectrum as arising from fluctuations in the halo counts and a residual scatter at fixed halo counts. In this Appendix, we show the conditional scatter of the tSZ power spectrum at fixed halo counts is Gaussian, with variance depending on the number of halos.

Consider a fixed halo catalog, specified by the halo counts $\{N_i\}$ in sufficiently fine mass–redshift bins. Once $\{N_i\}$ is fixed, the remaining randomness in the 1-halo tSZ power spectrum comes from the angular configuration of halos on the sky. Since the tSZ field is a superposition of extended halo profiles, different angular configurations lead to different amounts of profile overlap and hence to a residual scatter in the measured power spectrum.

We expand the Compton- y field in spherical harmonics,

$$y(\hat{\mathbf{n}}) = \sum_{\ell m} y_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}), \quad (\text{A1})$$

with

$$y_{\ell m} = \sum_a \tilde{y}_{\ell,a} Y_{\ell m}^*(\hat{\mathbf{n}}_a), \quad (\text{A2})$$

where $\tilde{y}_{\ell,a}$ reduces to the two-dimensional Fourier transform of the profile of halo a in the flat-sky limit, and $\hat{\mathbf{n}}_a$ is its sky position. The power-spectrum estimator is

$$\hat{C}_\ell = \frac{1}{2\ell+1} \sum_m |y_{\ell m}|^2. \quad (\text{A3})$$

Using the addition theorem, this can be written as

$$\hat{C}_\ell = \frac{1}{4\pi} \sum_a y_{\ell,a}^2 + \frac{1}{4\pi} \sum_{a \neq b} y_{\ell,a} y_{\ell,b} P_\ell(\hat{\mathbf{n}}_a \cdot \hat{\mathbf{n}}_b), \quad (\text{A4})$$

where the P_ℓ are the Legendre polynomials. The first term is the self-contribution of each halo, while the second depends on the relative angular configuration of distinct halos. The latter therefore describes the additional scatter induced by the overlap of halo profiles at fixed halo catalog.

Averaging only over the halo positions while keeping the catalog fixed, statistical isotropy implies

$$\langle P_\ell(\hat{\mathbf{n}}_a \cdot \hat{\mathbf{n}}_b) \rangle_{\{N_i\}} = 0, \quad a \neq b \quad (\text{A5})$$

for $\ell > 0$, and hence

$$\langle \hat{C}_\ell \rangle_{\{N_i\}} = \frac{1}{4\pi} \sum_a y_{\ell,a}^2 \equiv \bar{C}_\ell(\{N_i\}). \quad (\text{A6})$$

In binning notation, this becomes

$$\bar{C}_\ell(\{N_i\}) = \frac{1}{\Omega_{\text{sky}}} \sum_i N_i S_i, \quad (\text{A7})$$

where $S_i = |\tilde{y}_\ell(M_i, z_i)|^2$.

The fluctuation about this conditional mean is therefore

$$\delta \hat{C}_\ell \equiv \hat{C}_\ell - \bar{C}_\ell(\{N_i\}) = \frac{1}{4\pi} \sum_{a \neq b} y_{\ell,a} y_{\ell,b} P_\ell(\hat{\mathbf{n}}_a \cdot \hat{\mathbf{n}}_b), \quad (\text{A8})$$

so the residual scatter at fixed halo counts is sourced by the off-diagonal pair terms. The conditional covariance of the $\delta \hat{C}_\ell$ is

$$\begin{aligned} \text{cov}(\delta \hat{C}_\ell, \delta \hat{C}_{\ell'})_{\{N_i\}} &= \frac{1}{(4\pi)^2} \sum_{a \neq b} \sum_{c \neq d} \tilde{y}_{\ell,a} \tilde{y}_{\ell,b} \tilde{y}_{\ell',c} \tilde{y}_{\ell',d} \\ &\quad \times \langle P_\ell(\hat{\mathbf{n}}_a \cdot \hat{\mathbf{n}}_b) P_{\ell'}(\hat{\mathbf{n}}_c \cdot \hat{\mathbf{n}}_d) \rangle_{\{N_i\}}. \end{aligned} \quad (\text{A9})$$

The average over halo locations vanishes for $\ell \neq 0$ and $\ell' \neq 0$ unless the sum over four halos reduces to two pairs, i.e., $a = c$ and $b = d$ or $a = d$ and $b = c$, which contribute equally. Using the addition theorem, we have

$$\langle P_\ell(\hat{\mathbf{n}}_a \cdot \hat{\mathbf{n}}_b) P_{\ell'}(\hat{\mathbf{n}}_a \cdot \hat{\mathbf{n}}_b) \rangle_{\{N_i\}} = \frac{1}{2\ell+1} \delta_{\ell\ell'}, \quad a \neq b, \quad (\text{A10})$$

so that

$$\text{cov}(\delta\hat{C}_\ell, \delta\hat{C}_{\ell'})_{\{N_i\}} = \frac{2\delta_{\ell\ell'}}{2\ell+1} \frac{1}{(4\pi)^2} \sum_{a \neq b} \tilde{y}_{\ell,a}^2 \tilde{y}_{\ell,b}^2. \quad (\text{A11})$$

For a large number of clusters, the sum of the diagonal $a = b$ terms that are excluded from the covariance in Eq. (A11) is small in comparison to those terms retained. Adding back in the diagonal terms, we can therefore approximate

$$\text{cov}(\delta\hat{C}_\ell, \delta\hat{C}_{\ell'})_{\{N_i\}} \approx \frac{2\delta_{\ell\ell'}}{2\ell+1} \bar{C}_\ell^2(\{N_i\}). \quad (\text{A12})$$

This is simply the disconnected (Gaussian) variance for the estimated power spectrum of a field with true power spectrum $\bar{C}_\ell(\{N_i\})$. If we further approximate the conditional fluctuations $\delta\hat{C}_\ell$ as being Gaussian distributed, the conditional bandpower distribution may be approximated as

$$P(C|\{N_i\}) \approx \frac{1}{\sqrt{2\pi\sigma_G^2(\{N_i\})}} \exp\left[-\frac{(C - \bar{C}(\{N_i\}))^2}{2\sigma_G^2(\{N_i\})}\right], \quad (\text{A13})$$

where, after multipole binning and accounting for partial sky coverage, the bandpower variance is

$$\sigma_G^2(\{N_i\}) \approx \frac{2\bar{C}_\ell^2(\{N_i\})}{(2\ell+1)\Delta\ell f_{\text{sky}}}. \quad (\text{A14})$$

Equation (A14) shows that the conditional Gaussian variance is, strictly speaking, a function of the halo catalog. In the main text, we neglect this mild dependence and evaluate it at the mean catalog, or equivalently at a typical power-spectrum amplitude. In particular, if we return to the exact expression (A11) and average over the halo catalog, we have

$$\text{cov}(\delta\hat{C}_\ell, \delta\hat{C}_{\ell'}) = \frac{2\delta_{\ell\ell'}}{2\ell+1} (C_\ell^{\text{tSZ,1h}})^2. \quad (\text{A15})$$

This follows from the rearrangement

$$\begin{aligned} \left\langle \sum_{a \neq b} \tilde{y}_{\ell,a}^2 \tilde{y}_{\ell,b}^2 \right\rangle &= \left\langle \sum_{ab} \tilde{y}_{\ell,a}^2 \tilde{y}_{\ell,b}^2 \right\rangle - \left\langle \sum_a \tilde{y}_{\ell,a}^4 \right\rangle \\ &\rightarrow \left\langle \sum_{ij} N_i N_j S_i S_j \right\rangle - \left\langle \sum_i N_i S_i^2 \right\rangle \\ &= \sum_{ij} \lambda_i \lambda_j S_i S_j \\ &= \left(C_\ell^{\text{tSZ,1h}} \right)^2, \end{aligned} \quad (\text{A16})$$

where $\lambda_i = \langle N_i \rangle$ is the mean number of halos in mass and redshift bin i , and we have assumed Poisson statistics for the $\{N_i\}$. Equation (A15) is simply the disconnected (Gaussian) covariance for a field with power spectrum

$C_\ell^{\text{tSZ,1h}}$. The full covariance of $\hat{C}_\ell$ has an additional contribution from $\text{cov}(\bar{C}_\ell(\{N_i\}), \bar{C}_{\ell'}(\{N_i\}))$, which is exactly the trispectrum term:

$$\begin{aligned} \text{cov}(\bar{C}_\ell(\{N_i\}), \bar{C}_{\ell'}(\{N_i\})) &= \frac{1}{(4\pi)^2} \sum_i \lambda_i |\tilde{y}_\ell(M_i, z_i)|^2 \\ &\quad \times |\tilde{y}_{\ell'}(M_i, z_i)|^2. \end{aligned} \quad (\text{A17})$$

In the *Planck*-like regime considered in this paper, the dominant non-Gaussianity of the tSZ power spectrum still arises from fluctuations in the abundance of rare massive halos, while the additional scatter at fixed halo counts is subdominant. A more complete derivation in the halo-model framework is given by Ref. [61].

APPENDIX B: DETAILS OF SKY MAP GENERATION

1. Generating cluster catalogs

We assess the agreement between our synthetic cluster catalogs and the corresponding theoretical predictions. The catalogs are constructed with a minimum halo mass of $M_{\min} = 10^{14} M_\odot$ and a maximum mass of $M_{\max} = 10^{16} M_\odot$, spanning a redshift range of $z_{\min} = 0.005$ to $z_{\max} = 3.0$. Figure 15 shows the difference between the theoretical prediction and the mean cluster number counts as functions of redshift and mass, computed by averaging over 100 synthetic catalogs using precision parameters $n_M = 5000$ and $n_z = 5000$, where n_M and n_z are the number of mass and redshift bins we used to compute HMF. Since the tSZ power spectrum is particularly sensitive to the abundance of low-redshift local clusters, this choice of precision parameters ensures that the catalog generation correctly reproduces the expected cluster abundance in the regime most relevant for the power-spectrum analysis. The synthetic catalogs show good agreement with the theoretical predictions. For the number counts binned in redshift, 69.1% lie within the expected $\pm 1\sigma$ Poisson uncertainty, while for the counts binned in mass, 76.8% fall within the same range. These coverage statistics indicate that the synthetic catalogs are consistent with the theoretical predictions in cluster number counts.

2. Painting clusters onto sky maps

Once the catalogs are generated, each cluster is assigned a beam-convolved Compton- y profile based on its mass, redshift, and assumed pressure profile. The clusters are then projected onto a full-sky HEALPix grid to produce the synthetic y -maps. We additionally implement the GNFW pressure profile, parameterized by Arnaud *et al.* [31], on top of the default pressure profile, given by Battaglia *et al.* [62], in XGPaint. In general, the pressure profile, $y(M, z, \theta)$, depends on two parameters, M and z , and the angular coordinate θ . One might build a three-dimensional interpolation grid to compute the pressure profile at any

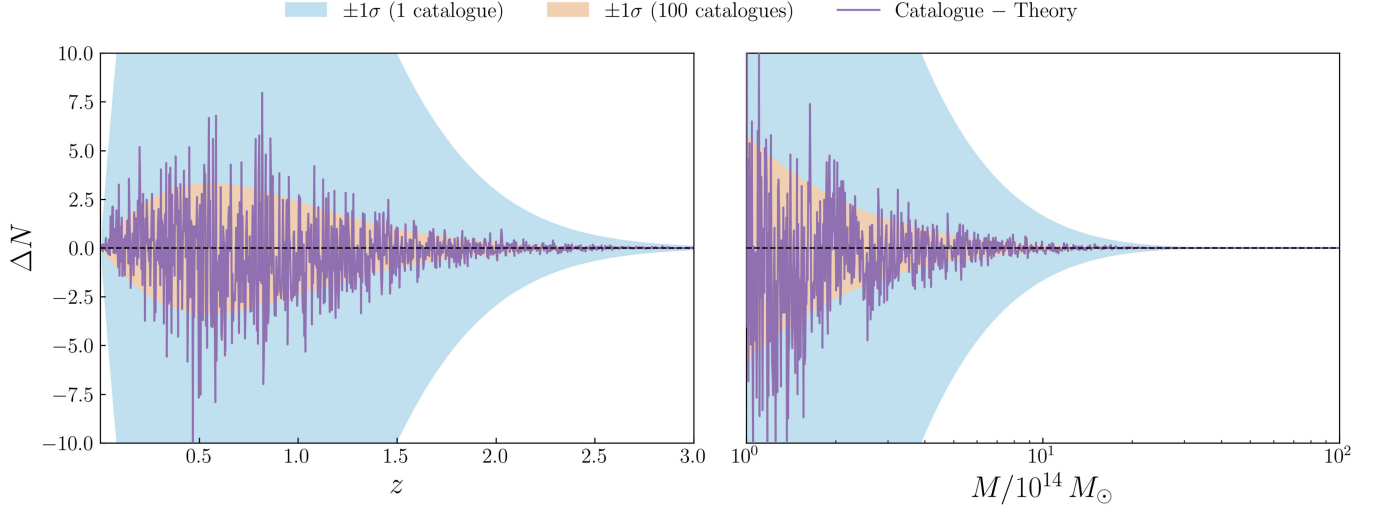


FIG. 15. Residuals between the mean number counts and the theoretical prediction (purple), shown with Poisson error bars for a single catalog (blue) and for 100 catalogs (orange). From the coverage statistics, the synthetic catalogs agree well with the theoretical predictions.

given mass, redshift and angle. This would need to be reconstructed every time the cosmological parameters are varied. However, the GNFW pressure profile with a constant concentration parameter, c_{500} , reduces this to an effectively one-dimensional problem. Specifically, the GNFW pressure profile can be rewritten as $y(M, z, \theta) = y_0(M, z)y_t(\theta/\theta_{500})$, where y_0 is the amplitude of the Compton- y parameter evaluated at the cluster center, and y_t is the spatial template of the cluster. In our case, y_0 and y_t can be analytically expressed as

$$y_0 = \frac{2\sigma_T}{m_e c^2} R_{500} P_{500} I\left(\frac{\theta}{\theta_{500}} = 0\right), \quad (\text{B1})$$

where

$$P_{500} = 1.65 \left(\frac{H(z)}{H_0}\right)^{8/3} P_0 \left(\frac{h}{0.7}\right)^2 \times \left[\frac{M_{500}}{3 \times 10^{14} h_{70}^{-1} M_\odot}\right]^{2/3+0.12} \text{eV cm}^{-3} \quad (\text{B2})$$

and

$$I\left(\frac{\theta}{\theta_{500}}\right) = \int_0^\infty du \left[c_{500} \sqrt{u^2 + \left(\frac{\theta}{\theta_{500}}\right)^2} \right]^{-\gamma} \times \left\{ 1 + \left[c_{500} \sqrt{u^2 + \left(\frac{\theta}{\theta_{500}}\right)^2} \right]^\alpha \right\}^{(\gamma-\beta)/\alpha} \quad (\text{B3})$$

and $y_t = I(\theta/\theta_{500})/I(0)$. To include beam convolution, we construct a two-dimensional interpolation grid in θ and θ_{500} that is fast and more accurate with high-precision parameters of $N_{\log \theta} = N_{\log \theta_{500}} = 8192$. To perform the beam convolution, we use the Fast Hankel Transform method in `Pixell.jl` to convolve the profiles with the beam with the same precision parameter on the θ -grid. We choose the real-space angular range to be $\log \theta_{\min} = -16.5$ and $\log \theta_{\max} = 2.5$, which corresponds to an ℓ range of $[0.082, 1.5 \times 10^7]$.

-
- [1] N. Clerc, M. Pierre, F. Pacaud, and T. Sadibekova, The cosmological analysis of X-ray cluster surveys—I. A new method for interpreting number counts: The CR-HR method for X-ray cluster surveys, *Mon. Not. R. Astron. Soc. U.S.A.* **423**, 3545 (2012).
- [2] T. M. C. Abbott *et al.* (DES Collaboration), Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing, *Phys. Rev. D* **98**, 043526 (2018).

- [3] M. Asgari *et al.* (KiDS Collaboration), KiDS-1000 Cosmology: Cosmic shear constraints and comparison between two point statistics, *Astron. Astrophys.* **645**, A104 (2021).
- [4] R. Mandelbaum *et al.* (LSST Dark Energy Science Collaboration), The LSST Dark Energy Science Collaboration (DESC) science requirements document, [arXiv:1809.01669](https://arxiv.org/abs/1809.01669).
- [5] P. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi,

- A. Banday, R. Barreiro *et al.*, Planck 2013 results. XXI. All-sky Compton parameter power spectrum and high-order statistics, *Astron. Astrophys.* **571**, A21 (2014).
- [6] L. Salvati, M. Douspis, and N. Aghanim, Constraints from thermal Sunyaev-Zel'dovich cluster counts and power spectrum combined with CMB, *Astron. Astrophys.* **614**, A13 (2018).
- [7] E. M. George *et al.*, A measurement of secondary cosmic microwave background anisotropies from the 2500-square-degree SPT-SZ survey, *Astrophys. J.* **799**, 177 (2015).
- [8] Í. Zubeldia, J.-B. Melin, J. Chluba, and R. Battye, The Planck SZiFi catalogues: A new set of Planck catalogues of Sunyaev-Zeldovich-detected galaxy clusters, *arXiv:2408.06189*.
- [9] R. A. Sunyaev and Y. B. Zeldovich, The observations of relic radiation as a test of the nature of X-ray radiation from the clusters of galaxies, *Comments Astrophys. Space Phys.* **4**, 173 (1972).
- [10] A. Sabyr, J. C. Hill, and Z. Haiman, Constraining cosmology with thermal Sunyaev-Zel'dovich maps: Minkowski functionals, peaks, minima, and moments, *Phys. Rev. D* **111**, 103536 (2025).
- [11] P. A. R. Ade *et al.*, Planck 2015 results XXIV, Planck 2015 results. XXIV. Cosmology from Sunyaev-Zeldovich cluster counts, *Astron. Astrophys.* **594**, A24 (2016).
- [12] Í. Zubeldia and A. Challinor, Cosmological constraints from Planck galaxy clusters with CMB lensing mass bias calibration, *Mon. Not. R. Astron. Soc.* **489**, 401 (2019).
- [13] S. Bocquet, S. Grandis, L. E. Bleem, M. Klein, J. J. Mohr, Schrabback *et al.*, SPT clusters with DES and HST weak lensing. II. Cosmological constraints from the abundance of massive halos, *Phys. Rev. D* **110**, 083510 (2024).
- [14] I. n. Zubeldia, B. Bolliet, A. Challinor, and W. Handley, Extracting cosmological information from the abundance of galaxy clusters with simulation-based inference, *Phys. Rev. D* **112**, 083536 (2025).
- [15] E. Komatsu and T. Kitayama, Sunyaev-Zeldovich fluctuations from spatial correlations between clusters of galaxies, *Astrophys. J.* **526**, L1 (1999).
- [16] E. Komatsu and U. Seljak, The Sunyaev-Zel'dovich angular power spectrum as a probe of cosmological parameters, *Mon. Not. R. Astron. Soc.* **336**, 1256 (2002).
- [17] I. G. McCarthy, A. M. C. Le Brun, J. Schaye, and G. P. Holder, The thermal Sunyaev-Zel'dovich effect power spectrum in light of Planck, *Mon. Not. R. Astron. Soc.* **440**, 3645 (2014).
- [18] A. Rotti, B. Bolliet, J. Chluba, and M. Remazeilles, Removing the giants and learning from the crowd: A new SZ power spectrum method and revised Compton y -map analysis, *Mon. Not. R. Astron. Soc.* **503**, 5310 (2021).
- [19] Planck Collaboration (Planck Collaboration), Planck 2015 results. XXII. A map of the thermal Sunyaev-Zeldovich effect, *Astron. Astrophys.* **594**, A22 (2015).
- [20] B. Bolliet, B. Comis, E. Komatsu, and J. F. Macías-Pérez, Dark energy constraints from the thermal Sunyaev-Zeldovich power spectrum, *Mon. Not. R. Astron. Soc.* **477**, 4957 (2018).
- [21] H. Tanimura, M. Douspis, N. Aghanim, and L. Salvati, Constraining cosmology with a new all-sky Compton parameter map from the Planck PR4 data, *Mon. Not. R. Astron. Soc.* **509**, 300 (2021).
- [22] S. Raghunathan *et al.* (SPTpol Collaboration and SPT-3G Collaboration), Measurement of the full shape of the thermal Sunyaev-Zel'dovich power spectrum from the south pole telescope and Herschel-SPIRE observations, *Astrophys. J. Lett.* **1002**, L22 (2026).
- [23] P. Zhang and R. K. Sheth, The probability distribution function of the SZ power spectrum: An analytical approach, *Astrophys. J.* **671**, 14 (2007).
- [24] M. W. Peel, R. A. Battye, and S. T. Kay, Statistics of the Sunyaev-Zel'dovich effect power spectrum, *Mon. Not. R. Astron. Soc.* **397**, 2189 (2009).
- [25] K. Cranmer, J. Brehmer, and G. Louppe, The frontier of simulation-based inference, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
- [26] G. Papamakarios and I. Murray, Fast ϵ -free inference of simulation models with Bayesian conditional density estimation, *arXiv:1605.06376*.
- [27] G. Papamakarios, D. C. Sterratt, and I. Murray, Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, *arXiv:1805.07226*.
- [28] D. J. Fixsen, The temperature of the cosmic microwave background, *Astrophys. J.* **707**, 916 (2009).
- [29] A. Cooray and R. K. Sheth, Halo models of large scale structure, *Phys. Rep.* **372**, 1 (2002).
- [30] J. L. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. S. Warren, G. Yepes, S. Gottlober, and D. E. Holz, Toward a halo mass function for precision cosmology: The limits of universality, *Astrophys. J.* **688**, 709 (2008).
- [31] M. Arnaud, G. W. Pratt, R. Piffaretti, H. Böhringer, J. H. Croston, and E. Pointecouteau, The universal galaxy cluster pressure profile from a representative sample of nearby systems (rexcess) and the rsz - $m500$ relation, *Astron. Astrophys.* **517**, A92 (2010).
- [32] E. Rasia *et al.*, Temperature structure of the intracluster medium from smoothed-particle hydrodynamics and adaptive-mesh refinement simulations, *Astrophys. J.* **791**, 96 (2014).
- [33] R. E. Angulo, V. Springel, S. D. M. White, A. Jenkins, C. M. Baugh, and C. S. Frenk, Scaling relations for galaxy clusters in the Millennium-XXL simulation, *Mon. Not. R. Astron. Soc.* **426**, 2046 (2012).
- [34] G. E. Addison, J. Dunkley, and D. N. Spergel, Modelling the correlation between the thermal Sunyaev Zel'dovich effect and the cosmic infrared background, *Mon. Not. R. Astron. Soc.* **427**, 1741 (2012).
- [35] Í. Zubeldia, J.-B. Melin, J. Chluba, and R. Battye, The Planck SZiFi catalogues: A new set of Planck catalogues of Sunyaev-Zeldovich-detected galaxy clusters, *Mon. Not. R. Astron. Soc.* **539**, 185 (2025).
- [36] 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. R. Astron. Soc.* **511**, 1771 (2022).
- [37] B. Bolliet, A. S. Mancini, J. C. Hill, M. Madhavacheril, H. T. Jense, E. Calabrese, and J. Dunkley, High-accuracy emulators for observables in Λ CDM, n_{eff} , σ_{m_ν} , and w cosmologies, *Mon. Not. R. Astron. Soc.* **531**, 1351 (2024).

- [38] B. Bolliet *et al.*, Class_sz I: Overview, *EPJ Web Conf.* **293**, 00008 (2024).
- [39] B. Bolliet *et al.*, CLASS_SZ II: Notes and examples of fast and accurate calculations of Halo model, large scale structure and cosmic microwave background observables, [arXiv:2507.07346](#).
- [40] J. Bradbury, R. Frostig, P. Hawkins, M.J. Johnson, C. Leary, D. Maclaurin, G. Neca, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable transformations of Python + NumPy programs (2018).
- [41] I. n. Zubeldia and B. Bolliet, cosmocnc: A fast, flexible, and accurate framework for galaxy cluster number count likelihood computation, *J. Cosmol. Astropart. Phys.* **11** (2024) 018.
- [42] M. Tomasi and Z. Li, Healpix.jl: Julia-only port of the HEALPix library (2021), [ascl:2109.028](#).
- [43] J. Boelts *et al.*, sbi reloaded: A toolkit for simulation-based inference workflows, *J. Open Source Software* **10**, 7754 (2025).
- [44] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, [arXiv:1711.05101](#).
- [45] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, [arXiv:1705.07057](#).
- [46] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, [arXiv:1906.04032](#).
- [47] L. Biewald, Experiment tracking with weights and biases (2020), software available from wandb.com.
- [48] J. Alvey, C. R. Contaldi, and M. Pieroni, Simulation-based inference with deep ensembles: Evaluating calibration uncertainty and detecting model misspecification, *Mach. Learn. Sci. Tech.* **7**, 015008 (2026).
- [49] J. Torrado and A. Lewis, Cobaya: Code for Bayesian analysis of hierarchical physical models, *J. Cosmol. Astropart. Phys.* **05** (2021) 057.
- [50] J. Torrado and A. Lewis, Cobaya: Bayesian analysis in cosmology (2019), [ascl:1910.019](#).
- [51] A. Gelman and D. B. Rubin, Inference from iterative simulation using multiple sequences, *Stat. Sci.* **7**, 457 (1992).
- [52] 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, [arXiv:2110.06581](#).
- [53] I. G. McCarthy *et al.*, The FLAMINGO project: Revisiting the S8 tension and the role of Baryonic physics, *Mon. Not. R. Astron. Soc.* **526**, 5494 (2023).
- [54] G. Efstathiou and F. McCarthy, The power spectrum of the thermal Sunyaev-Zeldovich effect, *Mon. Not. R. Astron. Soc.* **540**, 1055 (2025).
- [55] L. D. Shaw, O. Zahn, G. P. Holder, and O. Doré, Sharpening the precision of the Sunyaev-Zel'dovich power spectrum, *Astrophys. J.* **702**, 368 (2009).
- [56] K. Osato and M. Takada, Super sample covariance of the thermal Sunyaev-Zel'dovich effect, *Phys. Rev. D* **103**, 063501 (2021).
- [57] P. Ade *et al.* (Simons Observatory Collaboration), The Simons Observatory: Science goals and forecasts, *J. Cosmol. Astropart. Phys.* **02** (2019) 056.
- [58] A. Lewis, GetDist: A Python package for analysing Monte Carlo samples, *J. Cosmol. Astropart. Phys.* **08** (2025) 025.
- [59] www.csd3.cam.ac.uk.
- [60] www.dirac.ac.uk.
- [61] Z. Atkins, Certain Uncertainties: Methods for Accurate Error Estimation in the ACT DR6 Cosmology Results, Ph.D. thesis, Princeton University, 2025.
- [62] N. Battaglia, J. R. Bond, C. Pfrommer, and J. L. Sievers, On the cluster physics of Sunyaev-Zel'dovich and X-ray surveys. II. Deconstructing the thermal SZ power spectrum, *Astrophys. J.* **758**, 75 (2012).