

From few to many maps: A fast map-level emulator for extreme augmentation of cosmic microwave background systematics datasets

P. Campeti^{1,2,*}, J.-M. Delouis³, L. Pagano^{1,4,5}, E. Allys⁶, M. Lattanzi¹, and M. Gerbino^{1,4}

¹ INFN Sezione di Ferrara, Via Saragat 1, 44122 Ferrara, Italy

² ICSC, Centro Nazionale “High Performance Computing, Big Data and Quantum Computing”, Casalecchio di Reno, Italy

³ Laboratoire d’Océanographie Physique et Spatiale (LOPS), Univ. Brest, CNRS, Ifremer, IRD, 29200 Brest, France

⁴ Dipartimento di Fisica e Scienze della Terra, Università degli Studi di Ferrara, via Saragat 1, 44122 Ferrara, Italy

⁵ Institut d’Astrophysique Spatiale, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Bât. 121, 91405 Orsay Cedex, France

⁶ Laboratoire de Physique de l’Ecole Normale Supérieure, ENS, Univ. PSL, CNRS, Sorbonne Univ., Univ. Paris Cité, 75005 Paris, France

Received 14 March 2025 / Accepted 10 June 2025

ABSTRACT

Context. Generating massive sets of end-to-end simulations of time-ordered data for Monte Carlo analyses in cosmic microwave background (CMB) experiments typically incurs exceedingly high computational costs.

Aims. To address this challenge, we introduce a novel, fast, and efficient generative model built upon scattering covariances, the most recent iteration of the scattering transform statistics. This model is designed to augment by several orders of magnitude the number of map simulations in datasets of computationally expensive CMB instrumental systematics simulations, including their non-Gaussian and inhomogeneous features. Unlike conventional neural network-based algorithms, this generative model requires only a minimal number of training samples, making it highly compatible with the computational constraints of typical CMB simulation campaigns. While our primary focus is on spherical data, the framework is inherently versatile and readily applicable to 1D and 2D planar data, leveraging the localized nature of scattering statistics.

Methods. We validated the method using realistic simulations of CMB systematics, which are particularly challenging to emulate, and performed extensive statistical tests to confirm its ability to produce new statistically independent approximate realizations.

Results. Remarkably, even when trained on as few as ten simulations, the emulator closely reproduces key summary statistics – including the angular power spectrum, scattering coefficients, and Minkowski functionals – and provides pixel covariance estimates with substantially reduced sample noise compared to those obtained without augmentation.

Conclusions. The proposed approach has the potential to shift the paradigm in simulation campaign design. Instead of producing large numbers of low- or medium-accuracy simulations, future pipelines can focus on generating a few high-accuracy simulations that are then efficiently augmented using such a generative model. This promises significant benefits not only for current and forthcoming cosmological surveys such as *Planck*, *LiteBIRD*, Simons Observatory, CMB-S4, Euclid, and Rubin-LSST, but also for diverse fields including oceanography and climate science. We make both the general framework for scattering transform statistics, *HealpixML*, and the emulator, *CMBSAT*, available to the community.

Key words. methods: data analysis – methods: statistical – cosmic background radiation

1. Introduction

Cosmic microwave background (CMB) experiments face the significant challenge of generating massive sets of end-to-end simulations of time-ordered data (TOD) for Monte Carlo analyses. These simulations incur exceedingly high computational costs, often surpassing $O(100)$ million CPU hours (Planck Collaboration XII 2016). Despite their complexity and expense, these datasets are essential for accurately accounting for all relevant instrumental systematic effects in the statistical inference of cosmological parameters. This process is crucial at both the experiment-design and data-exploitation stages, as is exemplified by missions such as the *Planck* satellite (launched in 2009, Planck Collaboration VI 2020) and the upcoming *LiteBIRD* satellite (scheduled for launch in the early 2030s, LiteBIRD collaboration 2023).

The computational demands vary according to the specific experiment, as well as factors such as the scanning strategy, the number of detectors simulated, the chosen simulation framework, and the systematic effects included. Beyond CMB experiments, similar challenges also arise in other forthcoming cosmological surveys, such as Rubin-LSST and Euclid. In non-cosmological fields, particularly in climate and ocean sciences, simulating spherical datasets can be even more computationally intensive (see e.g., Acosta et al. 2024 for the climate Coupled Model Intercomparison Project in its sixth phase (CMIP6) effort), with some applications requiring up to $O(10^3)$ million CPU hours.

Due to their high computational costs, only a limited number of end-to-end simulations – typically a few hundred – are available for cosmological inference. For example, the most recent *Planck* data release included just 400 simulated maps (Planck collaboration LVII 2020; Tristram et al. 2022) with

* Corresponding author.

simplified instrumental errors to make it computable. However, recent studies (Beck et al. 2022) suggest that this number may be insufficient for future CMB experiments. Achieving high-accuracy estimation of cosmological parameters, such as the tensor-to-scalar ratio, may for instance require at least $O(10^{4-5})$ simulations. This would correspond to an overwhelming computational cost of approximately $O(10^{4-6})$ million CPU hours, far exceeding the typical resource allocations for such scientific projects. Furthermore, machine learning applications to CMB data, such as direct cosmological parameter extraction from CMB maps using simulation-based inference (Wolz et al. 2023), would demand even larger simulation sets ($\sim O(10^{5-6})$ samples) to construct suitable training datasets.

To address these challenges, we propose a novel approach to dramatically reduce the computational cost of simulation campaigns by augmenting a dataset of a few end-to-end map simulations. We show that this data augmentation can be achieved with maximum entropy generative models built from scattering transform (ST) statistics. Such models allow one to generate an approximate realization of a process conditioned on an intermediate low-dimensional (i.e., latent) representation of the target field statistics (Bruna & Mallat 2019; Allys et al. 2020). Specifically, here we adopt as a latent representation the latest iteration of ST, the scattering covariance (hereafter SC Cheng et al. 2024). In contrast to other generative models purely based on neural networks – such as generative adversarial networks (GANs) (Goodfellow et al. 2014), variational auto-encoders (VAEs) (Kingma & Welling 2013), or diffusion models (Sohl-Dickstein et al. 2015) – with SC we can generate new samples from very small “training” datasets, or even from a single realization (i.e., “one-shot” or “few-shots learning,” see e.g. Jeffrey et al. 2022; Mousset et al. 2024). The concepts and tools related to these generative models also allow for the development of efficient component separation algorithms, which can be applied to a limited amount of data, and even directly on observations (Delouis et al. 2022; Auclair et al. 2024).

The SC (also known as scattering spectra, Cheng et al. 2024) represents the last iteration of the ST statistics, a family of very efficient and powerful summary statistics originally proposed in Mallat (2012); Bruna & Mallat (2013) to extract information from highly non-Gaussian signals. The general features of ST statistics are the use of convolution with fixed wavelet filters to separate a field into its individual scales, before evaluating the interaction between these scales using nonlinear operations, possibly in an iterative fashion. These statistics include wavelet scattering transforms (WST, Bruna & Mallat 2013; Allys et al. 2019), wavelet phase harmonics (WPH, Zhang & Mallat 2019; Allys et al. 2020), and SCs, or scattering spectra (Morel et al. 2022; Cheng et al. 2024). Crucially, ST requires no training because its wavelet kernels are fixed, enabling the augmentation of a small training set of tens or hundreds of end-to-end simulations. Moreover, because the wavelet kernels are not learned, the model is more interpretable and has generally fewer free parameters than, for instance, a convolutional neural network (CNN).

The ST statistics has proven effective across a wide range of applications (see e.g. Cheng & Ménard 2021, for a review) and recently made a breakthrough in astrophysics and cosmology, excelling in tasks such as parameter estimation (Cheng et al. 2020; Eickenberg et al. 2022; Valogiannis & Dvorkin 2022; Gatti et al. 2024; Valogiannis et al. 2024; Blancard et al. 2024), classification (Bruna & Mallat 2013), the statistical description of complex structures (Allys et al. 2019; Régaldo-Saint Blancard et al. 2020), and component separation (Delouis et al. 2022;

Régaldo-Saint Blancard et al. 2021; Auclair et al. 2024). Moreover, ST has provided a powerful latent representation, enabling the development of maximum entropy generative models that can efficiently sample new approximate realizations (synthetic data or “emulations”) of homogeneous¹ astrophysical and cosmological fields, even from a limited number of simulations (Allys et al. 2020; Jeffrey et al. 2022; Régaldo-Saint Blancard et al. 2023; Price et al. 2023; Mousset et al. 2024). The generation of new synthetic data is typically achieved through gradient descent methods, exploiting statistical mechanics ideas to perform sampling, such as the microcanonical gradient descent (Bruna & Mallat 2019; Häggbom et al. 2024). Combining ST with gradient descent sampling offers a highly promising approach to address the issues of computational efficiency in statistical inference presented above. It enables the production of approximate new samples, and therefore augmentation of small ensembles of end-to-end simulations at a fraction of the cost of a full simulation campaign. The ST framework has been successfully applied also to spherical datasets, showcasing its effectiveness in tasks such as component separation (Delouis et al. 2022) and generative modeling (Mousset et al. 2024), with the latter being the main focus of this paper.

In this work, we present a case study using a full-sky simulation dataset that realistically models the instrumental systematics encountered in CMB satellite experiments to demonstrate the augmentation procedure. We test how the size of the “training” dataset impacts the emulations produced by the augmentation process, assuming that only a very limited number of end-to-end simulations – a few tens to a few hundred – are available. We validate the generative model (hereafter referred to as the emulator) through an extensive set of visual and statistical tests, requiring the emulated dataset to capture both the true data features (i.e., to not be biased compared to the true one) and their variability (i.e., to have the same variance).

A fast and efficient map-level emulator offers a number of practical advantages in many aspects of data analysis in cosmology. Firstly, generating a sufficient number of simulations for cosmological inference can be a resource-intensive task, and often one must deal with a limited number of simulations, which can introduce significant sample variance, particularly in empirical covariance matrices. While various approaches have been proposed to mitigate this issue (e.g., Sellentin & Heavens 2016), recent studies indicate that certain effects remain difficult to capture without substantially more simulations (Beck et al. 2022). In this context, we explore how the augmentation framework can mitigate sample noise in pixel-pixel covariance matrices when the simulation budget is limited to only a few tens or hundreds of expensive end-to-end simulations. By leveraging this framework, current and future CMB surveys such as *Planck* or *LiteBIRD* could better exploit existing data and reduce the nonoptimal error bars that often arise from using too few simulations. Moreover, having the ability to generate realistic surrogates from small datasets could shift the current paradigm in many simulation campaigns, allowing one to focus both human and computational effort on producing a few highly realistic simulations, instead of many simplified ones. Beyond the obvious computational budget and environmental considerations, reducing simulation costs is also crucial for enabling faster prototyping of many design choices (e.g., in experiment optimization, foreground modeling,

¹ Throughout this work we call a field homogeneous when its statistical properties do not depend on position, i.e. they are assumed to be translation invariant; when this symmetry is broken we refer to the field as inhomogeneous.

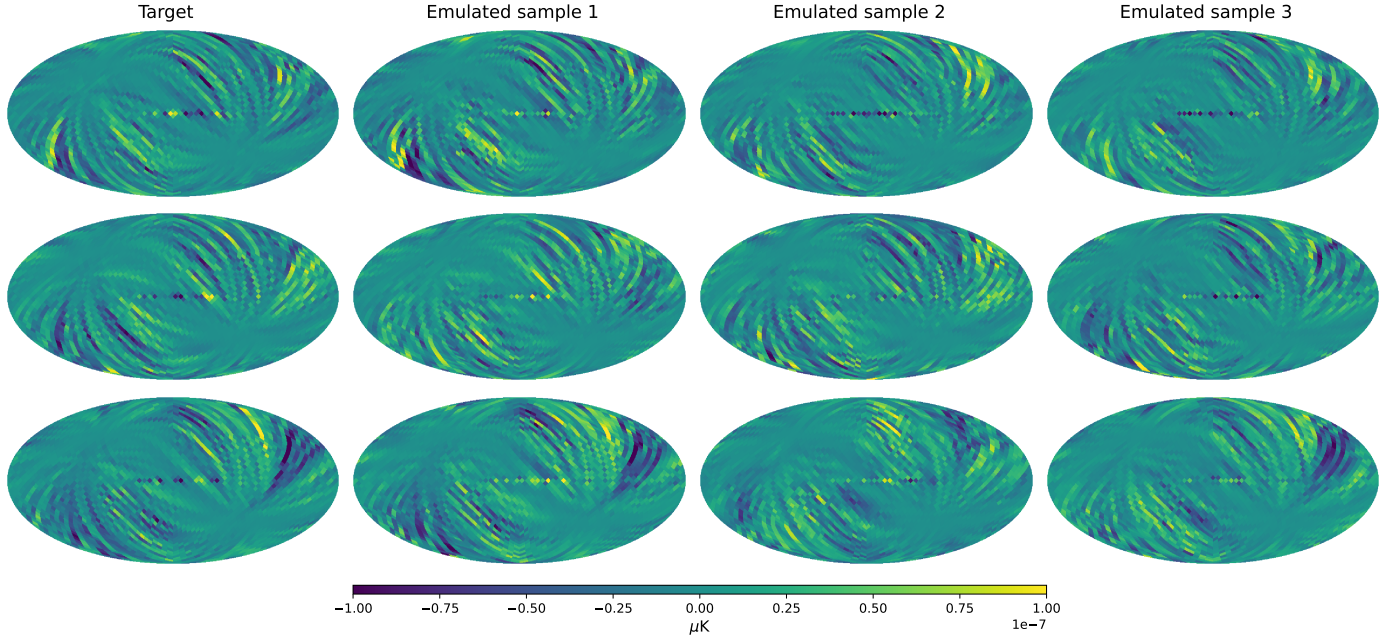


Fig. 1. Side-by-side comparison of an input Q map (from the moderate size input dataset, labeled “Target”, in Galactic coordinates) and three corresponding emulated samples synthesized from that input. The leftmost column displays the input map; the other columns present three distinct emulated samples obtained from our generative model, with each row showing a different input target map. The U map (not shown here) shows very similar behavior. See Sections 2 and 5.3 for further details.

and component separation). A fast emulator also has important applications in simulation-based inference (Cranmer et al. 2020) for cosmological parameters, which is crucial in case the likelihood is unknown or too expensive to evaluate, and/or complex systematics and foregrounds residuals are present in the maps and hard to characterize (Wolz et al. 2023).

While the main focus in this paper is applying the model to data on the sphere in HEALPix pixelization (Górski et al. 2005), we emphasize that our framework is inherently versatile and readily applicable to 2D planar and 1D data, leveraging the localized nature of the wavelet transform used to compute the ST statistics. Therefore, our framework can be directly applied to simulated or observed data from current and future CMB ground-based experiments; for instance, to flat-sky tiles analyzed by BICEP/Keck (BICEP collaboration 2021) or to partial-sky HEALPix-pixelized maps from the Small-Aperture Telescopes of Simons Observatory (SO) (Simons Observatory 2019) and CMB-S4 (Abazajian et al. 2016). By contrast, maps delivered in equirectangular (CAR) grids – such as the ones produced for SO’s Large-Aperture Telescope (LAT) survey – require extra preprocessing steps because the latitude-dependent pixel area of CAR grids can mimic genuine statistical inhomogeneity if left uncorrected.

The paper is organized as follows. In Section 2 we describe how we built the initial ensemble end-to-end of simulations. Section 3 offers a concise overview of the SC statistics, the generative model, and the data augmentation algorithm, along with a description of the software implementations. Section 4 provides an estimate of the computational cost of the emulator in various contexts and compares it to a typical full-sky simulation campaign (e.g., for *Planck* or *LiteBIRD*). In Section 5, we evaluate the performance of the emulator on two case studies: one trained on just 10 simulations and another on 100 ones. Finally, in Section 6, we summarize our findings and discuss future research directions.

2. Simulations

We demonstrate our methods on a very challenging scenario for SC emulation: highly inhomogeneous² maps arising from a variable gain systematic, a common instrumental artifact in satellite CMB experiments (e.g., *Planck*). In the following, we consider maps containing only this systematic effect, no white or other noises, no CMB signal or lensing, and no foregrounds.

We generated 10 000 end-to-end temperature and polarization Q/U simulations using the publicly available *litebird_sim*³ TOD simulation and map-making toolkit. Specifically, we simulated four detectors, arranged in two orthogonally polarized pairs, with a relative orientation of 45 degrees between the pairs. The detectors follow a *Planck*-like scanning strategy (Tauber et al. 2010), and we introduced random gain miscalibration by adding random Gaussian fluctuations at the TOD level. Each fluctuation was applied to chunks of 100 000 samples (~4 hours) with a standard deviation of 0.0001. The resulting maps were produced at a HEALPix resolution of $N_{\text{side}} = 128$, then downgraded to $N_{\text{side}} = 16$ to decrease computational resources consumption, since we do not expect our conclusions to vary significantly with resolution. The leftmost column of Figure 1 (labeled “Target”) shows three example realizations from these simulations.

The variable gain systematic, coupled with the scanning strategy, induces large-scale “stripe” features, mainly dominated by dipole leakage (similar to those shown in Delouis et al. 2019). Because the gain drift we inject depends on the scanning pattern, the resulting maps are strongly inhomogeneous, i.e., their pixel-to-pixel statistics vary across the sky, unlike a homogeneous CMB realization whose statistics depend only on angular separation. In this work, we validate the emulator on maps containing such highly inhomogeneous structures, which are challenging to

² See footnote 1 for definition of “inhomogeneous” in this context.

³ https://github.com/litebird/litebird_sim

emulate with the SC statistics. We discuss in more detail these challenges in Section 3. We stress, however, that the method is not limited to this specific systematic. Any effect that imprints spatially varying or non-Gaussian signatures at the map level (e.g., correlated $1/f$ noise, residual foregrounds after component separation, beam asymmetries, bandpass mismatch) manifests itself as structure in the SC space and could therefore be captured by the same procedure, even if the most complex processes may require some fine-tuning. The performance of the emulator is expected to improve (or at least not degrade) for systematics that are closer to homogeneous or Gaussian, because the optimization landscape becomes simpler.

3. Emulator

3.1. Basics of generative modelling

A generative model of a random vector, X , is an operator, $\widehat{G}$, that transforms a Gaussian white noise random vector, Z , into a model, $\widehat{X} = \widehat{G}(Z)$ of X (Ng & Jordan 2001). In other words, such a model is capable of producing new approximate synthetic observables that realistically approximate the true ones and preserve their underlying probability distribution.

In a broad sense, a simulation can also be viewed as a specific type of generative model in which the time evolution (i.e., dynamics) of a physical system is explicitly encoded, controlled by a prescribed set of parameters (see e.g. Price et al. 2023). Given a set of initial conditions, simulations evolve the system forward in time and return a synthetic observable at a desired point in this evolution. For cosmological observations such as CMB ones, a simulator (or simulation framework) takes an initial configuration of the observing instrument and produces a time series (TOD), i.e., a synthetic sky observation spanning a chosen duration.

By contrast, emulation is a generative model that creates an approximate mapping between the initial conditions and the resulting synthetic observables. In an end-to-end emulation approach, this mapping is learned directly and typically requires large training datasets, the size of which depends on the complexity and dimensionality of the system. When datasets are insufficient for direct end-to-end emulation, one can represent the data using a lower-dimensional summary statistic – often referred to as a latent representation – and then let the generative model learn the mapping from this reduced representation. Although some information is inevitably lost when using such a latent representation, the benefit is a greatly reduced need for extensive training data. The surjectivity of the latent representation introduces variability into the emulations, since several observables can in principle correspond to a single latent vector.

In the following, we identify two critical requirements for a successful emulator: predictiveness and representativeness. Predictiveness ensures that the emulator can reproduce true data features, for instance Galactic foregrounds and localized noise or systematics structures in the maps, or specific patterns in power spectra. Representativeness ensures instead that the emulator can produce (at least approximately) statistically independent realizations that are representative of the true underlying data distribution. Predictiveness and representativeness are tied to the concepts of bias and variance in statistical modeling, whereby good predictiveness is equivalent to reducing bias, while good representativeness links to capturing variance.

In cosmology, machine learning algorithms such as GANs (Goodfellow et al. 2014) and VAEs (Kingma & Welling 2013)

have been successfully employed for end-to-end emulation of relevant observables (Rodriguez et al. 2018; Chardin et al. 2019). However, these networks require large, expensive simulation datasets for initial training. Since the present work deals with relatively small to moderate training sets of CMB data – represented as a pixelized map (via the HEALPix scheme) – we instead leverage the SC statistics.

3.2. Scattering covariance (SC)

Scattering covariance (SC) is the latest iteration of the scattering transform (ST), originally developed by Mallat (2012) and Bruna & Mallat (2013) to provide a more interpretable framework for understanding CNNs (see e.g. Cheng & Ménard 2021, for a review). Initially applied to one-dimensional signals (Morel et al. 2022), SC has since been extended to two-dimensional planar fields (Cheng et al. 2024) and more recently generalized to data on the sphere (Mousset et al. 2024). SC operates by separating an input field, I , into different scales by convolving it with a family of complex-valued wavelet filters, Ψ_λ , where $\lambda = (j, \gamma)$ specifies both the scale, j , and orientation, γ . After each convolution, SCs are computed by applying a modulus to characterize the interaction between scales and, for higher-order coefficients, computing also covariances among moduli of wavelet-convolved fields to capture additional cross-scale couplings. In this work, we focus on four specific coefficients, used in Cheng et al. (2024); Mousset et al. (2024):

$$\begin{aligned} S_1^{\lambda_1} &= \langle |I * \Psi_{\lambda_1}| \rangle, \\ S_2^{\lambda_1} &= \langle |I * \Psi_{\lambda_1}|^2 \rangle, \\ S_3^{\lambda_1, \lambda_2} &= \text{Cov} \left[|I * \Psi_{\lambda_1}|, |I * \Psi_{\lambda_2}| * \Psi_{\lambda_1} \right], \\ S_4^{\lambda_1, \lambda_2, \lambda_3} &= \text{Cov} \left[|I * \Psi_{\lambda_3}| * \Psi_{\lambda_1}, |I * \Psi_{\lambda_2}| * \Psi_{\lambda_1} \right], \end{aligned} \quad (1)$$

where $*$ denotes convolution, covariances are defined as $\text{Cov}[X, Y] = \langle XY^* \rangle - \langle X \rangle \langle Y^* \rangle$ for two complex fields, X and Y , and $\langle \cdot \rangle$ indicates a spatial average. The first two coefficients measure the first- and second-order moments of single-scale responses, while the latter two encode two- and three-scale couplings, respectively. We also use in the following cross SC statistics computed on two different input fields; more details are reported in Appendix B.

In all of the results shown in this paper, we computed wavelet transforms by convolution with a family of dyadic⁴ wavelets in pixel space on the HEALPix sphere. The specific wavelet filters used in this work, together with details on our specific implementation, are reported in Appendix A. The range of scales probed by the wavelets is given by $J_{\min} \leq j \leq J_{\max}$ where $J_{\min} \geq 0$ and $J_{\max} = \log N_{\text{side}} / \log 2$, while the number of scales is given by $J = J_{\max} - J_{\min} + 1$. The angular resolution is given in terms of the number, R , of orientations. In this work, we adopt for the wavelet filters complex kernels with size 3×3 pixels, probing $J = 4$ different scales and $R = 4$ orientations.

As for the normalization of the SC coefficients in Eq. (1) – which is necessary to avoid issues while optimizing coefficients varying over several orders of magnitude – we adopted a different convention compared to Cheng et al. (2024) and Mousset et al. (2024), and normalized by the S_2 coefficient of the

⁴ The family of wavelets is built by dilating and rotating a mother wavelet. If the dilation factor in the scales is two, the wavelet family is named dyadic.

same field:

$$\bar{S}_1^{\lambda_1} = \frac{S_1^{\lambda_1}}{\sqrt{S_2^{\lambda_1}}}, \quad \bar{S}_3^{\lambda_1, \lambda_2} = \frac{S_3^{\lambda_1, \lambda_2}}{\sqrt{S_2^{\lambda_2} S_2^{\lambda_1}}}, \quad \bar{S}_4^{\lambda_1, \lambda_2, \lambda_3} = \frac{S_4^{\lambda_1, \lambda_2, \lambda_3}}{\sqrt{S_2^{\lambda_2} S_2^{\lambda_3}}},$$

while S_2 was left unnormalized. The total number of parameters in the SC coefficients amounts to 1472, subdivided into 16 parameters for S_1 and S_2 , 160 parameters for S_3 , and 1280 parameters for S_4 .

In principle, the SC model is designed for homogeneous fields, allowing one to estimate higher-order statistics even from a single observation (e.g., Jeffrey et al. 2022; Cheng et al. 2024). In our inhomogeneous scenario, additional information to catch the localized structures at large scales is required, which we find particularly efficient to supply by introducing orientation-informed wavelets: we define the SC statistics relative to some primary local orientation (in our case relative to the scanning strategy stripes), which is the orientation with the largest amplitude at each pixel and scale in the wavelet-convolved maps. This allows one to precompute a 4×4 (since in our case we deal with $R = 4$ wavelet orientations) rotation matrix for each pixel, which rotates the phase of wavelet-convolved maps (before taking the modulus) into the primary orientation basis. The SC statistics computed after this rotation will have amplitudes reflecting the alignment with the primary orientation. We describe the orientation-informed wavelets in more detail in Appendix A.

We also find that the sharp oscillations – especially at smaller scales – in the angular power spectra featured by the specific systematics we consider in this paper (see Fig. 3) are hard to reproduce correctly based only on the SC statistics. In fact, while the first two scattering coefficients can be interpreted as binned power spectra, the features we would like to resolve in the power spectra in this specific case are below the binning resolution introduced by the wavelet convolution. While in principle it is possible to solve this issue by changing the scaling of the wavelets to be non-dyadic, this introduces many additional parameters in the SC statistics. We find it very efficient and effective to solve this problem by instead adding a constraint directly on the angular power spectra of the maps (see Section 3.4 for further details).

3.3. Mean-field microcanonical gradient descent

We adopted a maximum-entropy microcanonical model as the generative framework, with approximate sampling performed via gradient descent. This model has been employed previously to address the one-shot and few-shots learning problem; namely, generating new approximations of a field given only a single or a few observations (Bruna & Mallat 2019).

The microcanonical framework allows for efficient sampling in high dimensions: sampling proceeds by transporting high-entropy initial states to the microcanonical ensemble, which approximate the ensemble of typical samples of the process under study. The sampling is performed in practice by initializing a map, x , as a random Gaussian white noise realization (whose variance matches the target homogeneous variance) and then performing a gradient descent (in pixel space in our specific case) under the loss function constraint,

$$\mathcal{L} = \|\Phi(x) - \Phi(\tilde{x})\|_2^2, \quad (2)$$

where Φ is the latent representation of choice (in our case, SC), x is the “running” map that is modified at each iteration of the

gradient descent, $\tilde{x}$ represents the target field, and $\|\cdot\|_2$ is the Euclidean ℓ_2 norm. The sampled map is the value of x at the end of the optimization.

However, the maximum-entropy microcanonical model is known to suffer from overfitting: entropy can collapse during gradient descent, causing the samples to be too similar to the target, limiting their variability. To mitigate this, one can employ mean-field gradient descent (Allys et al. 2020; Häggbom et al. 2024), which modifies the loss function (Eq. (2)) so that gradient descent acts simultaneously on a batch of m maps $\{x_j\}$ with $j = 1, \dots, m$ (initialized as independent initial white noise realizations), moving them collectively toward the target energy:

$$\mathcal{L} = \left\| \text{Ave}_j \Phi(x_j) - \Phi(\tilde{x}) \right\|_2^2, \quad (\text{mean-field gradient descent}), \quad (3)$$

where the average, Ave_j , is performed over the batch of samples, $\{x_j\}$, at each iteration. By moving the entire batch towards the target in aggregate, the mean-field gradient descent achieves a higher lower bound on entropy, thereby mitigating collapse. The final iteration of the simultaneous gradient descent gives in output a batch of m samples⁵. We adopted this strategy throughout the emulation procedure described below.

3.4. Augmentation

We now operate under the assumption that only a limited number of end-to-end simulations is available, a common scenario not only in cosmology but also in many other fields in which generating a large number of high-fidelity simulations may be prohibitively expensive. We therefore consider two representative sizes for the input dataset of simulations⁶: a small dataset of 10 samples and a moderate-sized dataset of 100 maps. Naturally, we expect that the size of this dataset will affect both the quality and diversity of the generated (emulated) samples, with the larger input dataset encoding more representative information on the underlying probability distribution of the data. We compare the generative model performance in these two cases in Section 5.

3.4.1. Single-target versus average-target

While Eq. (3) defines the mean-field gradient descent loss for generating new samples from just one target field, $\tilde{x}$, our approach can leverage an ensemble of $n > 1$ targets $\tilde{x}_i$ (with $i = 1, \dots, n$). We investigated two distinct strategies for defining the loss function incorporating multiple targets.

In the first strategy, which we name the average target approach, we used the average of the summary statistics from the input dataset as the target. The resulting loss function is defined as (Allys et al. 2020; Cheng et al. 2024)

$$\mathcal{L} = \left\| \text{Ave}_j \Phi(x_j) - \text{Ave}_i \Phi(\tilde{x}_i) \right\|_2^2. \quad (\text{average-target}) \quad (4)$$

This formulation ensures that the collective properties of the generated samples match the mean characteristics of the target ensemble.

⁵ The mean-field microcanonical gradient descent reduces to the conventional microcanonical gradient descent if the batch size is $m = 1$.

⁶ We name this “input” dataset rather than “training” set, since our generative approach based on the SC does not require standard training, unlike deep-learning, for instance.

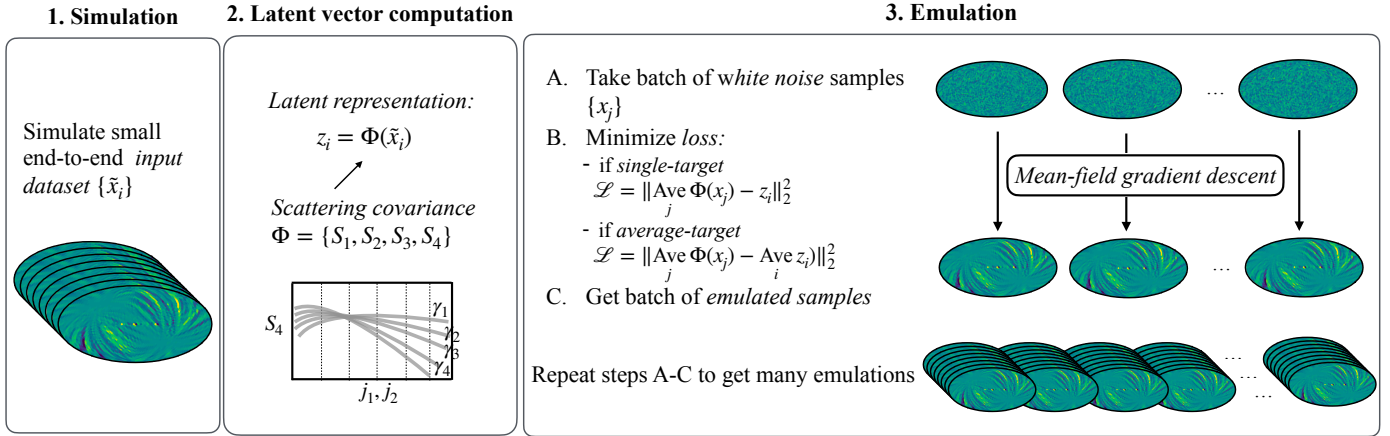


Fig. 2. Schematic view of the algorithm for extreme augmentation. See Section 3.4.2 for details.

The second strategy, which we name the single-target approach, transfers the input probability distribution into the emulated dataset by uniformly sampling a target from the input dataset (Price et al. 2023). In standard (i.e., non-mean-field) microcanonical gradient descent, this method prescribes randomly drawing a simulation, $\tilde{x}_i$, from the input dataset and then using it as the target in the loss Eq. (2). Since we used mean-field gradient descent, we adapted this prescription as follows. First we generated a batch of m new samples, $\{x_j\}$, for each of the n targets in the input dataset, minimizing

$$\mathcal{L} = \left\| \text{Ave}_j \Phi(x_j) - \Phi(\tilde{x}_i) \right\|_2^2. \quad (\text{single-target}) \quad (5)$$

Next, we uniformly sampled N targets with replacement from the input dataset, and selected, for each sampled target, an emulation from the corresponding batch. If the same target was sampled more than once, we ensured that each time a different emulated map was selected from the same batch. Thus, at the end, we would have a dataset of N emulations.

We applied the single- and the average-target strategies to both the small and the moderate-sized input datasets, finding that the optimal approach depends on the size of the input dataset. In the following, we always show results obtained with the average-target approach for the 10-input maps case and with the single-target approach for the 100-input maps one. We discuss this topic in more detail in Section 5.3.

3.4.2. Algorithm for extreme augmentation

We simultaneously minimized three different losses on the SC coefficients: two for the auto-SC coefficients (Eq. (1)) of the Q map and U map each and one on the cross-SC coefficients (i.e., computed cross-correlating the Q and U fields, see Appendix B), with the latter accounting for the statistical dependency between the two Stokes parameters.

As is explained in Section 3.2, we complemented the information encoded in the SC statistics by minimizing an additional loss on the $XX = EE, BB, EB$ angular power spectra of the maps:

$$\mathcal{L} = \sum_{\ell} \left\| \frac{\text{Ave}_j C_{\ell}^{XX,(p)}(x_j)}{\text{Std}_j C_{\ell}^{XX,(p)}(x_j)} - \frac{\text{Ave}_i C_{\ell}^{XX,(p)}(\tilde{x}_i)}{\text{Std}_i C_{\ell}^{XX,(p)}(\tilde{x}_i)} \right\|_2^2, \quad (6)$$

where

$$C_{\ell}^{XX,(p)} = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{\ell m}^{XX}|^p \quad (7)$$

is the angular power spectrum for the L^p -norm and Std indicates the standard deviation of the power spectrum. Specifically, we computed this loss at each step for both the L^1 - and the L^2 -norm power spectra and then summed both to the SC losses.

We illustrate the method for extreme augmentation in Figure 2. The augmentation process consists of three steps, with Steps 2 and 3 varying according to whether the average-target or single-target approach is chosen:

1. Simulation:

- Generate a small set of end-to-end simulations, forming the input dataset. In our specific example, these are HEALPix maps including only instrumental systematics (described in detail in Section 2).
- Normalize each map by subtracting the mean and dividing by the standard deviation of the input dataset to facilitate the optimization in Step 3 below, and revert this operation upon completion.

2. Latent vector computation:

- Compute the latent representation, $\Phi(\tilde{x}_i)$, of each $\tilde{x}_i$ in the input dataset. In our specific case, Φ is the SC statistics (see Section 3.2).

3. Emulation (i.e., sampling via gradient descent):

- Initialize a batch of m maps $\{x_j\}$ to random Gaussian white noise realizations.
- At each iteration, compute $\text{Ave}_j \Phi(x_j)$ and minimize in pixel space, simultaneously for all samples in the batch, the mean-field microcanonical loss in Eq. (3), where the target is either the latent representation, $\Phi(\tilde{x}_i)$, of one of the input maps (if we choose the single-target approach, Eq. (5)) or the average latent vector, $\text{Ave}_i \Phi(\tilde{x}_i)$ (i.e., the mean SC coefficients of all input maps, Eq. (4)) for the average-target approach.
- We minimize the sum of four different losses: the auto and cross SC coefficients of the Q map and U maps and the loss on the EE, BB , and EB angular power spectra of these maps (Eq. (6)).
- At the end of optimization, the resulting batch of m emulated maps will have latent representations close to their respective targets.

- Under the single-target approach, an additional step is required: one should uniformly sample N targets with replacement from the input dataset and, for each target, select one emulation from its corresponding batch.

Step 3 can be repeated as many times as necessary to produce an arbitrarily large number of emulated maps based on the input dataset.

While in standard microcanonical gradient descent running the optimization for too many steps can often cause entropy collapse, here mean-field gradient descent solves this issue. The optimization reaches convergence (i.e., a plateau in the loss) in all cases after a few hundred steps. We conservatively ran the minimization for at least 1000 steps, which was sufficient for the global loss to decrease by 4–5 orders of magnitude, resulting in subpercent accuracy on the SC coefficients.

4. Software

4.1. HealpixML and CMBS CAT packages

We computed SC coefficients on the sphere using the publicly available HealpixML software⁷, written in Python and TensorFlow. HealpixML implements – among other useful features, such as gradient descent computation and angular power spectrum implementation in TensorFlow – efficient pixel-space wavelet convolutions on the HEALPix grid using single or multiple GPUs (but can run as well on tradition CPUs), enabling a state-of-the-art performance for scattering analyses in spherical geometry. It supports SC calculation on 1D, 2D planar, and on the HEALPix sphere, and previous iterations of the code⁸ have been applied to *Planck* data for component separation purposes (Delouis et al. 2022). Benchmarks of HealpixML’s performance for this work are provided in Section 4.2. We note that several HealpixML features may be broadly useful to scientists in various fields. For instance, the fast and efficient real-space convolution routines can be repurposed in other machine learning applications, while the TensorFlow-based implementation of the HEALPix angular power spectrum computation routine (i.e., anafast) can be incorporated into fast likelihood analyses written in TensorFlow.

In addition, we provide the CMBS CAT package⁹, a pip-installable and user-friendly Python/TensorFlow implementation of the augmentation algorithm presented in this paper for emulating and dataset augmentation for CMB systematics. CMBS CAT heavily relies on the fast, efficient routines in HealpixML and can be readily adapted for various CMB systematics simulations or related applications. We also provide a tutorial jupyter notebook explaining all the features of the CMBS CAT package¹⁰. The optimization step described in Section 3.4 was performed using the L-BFGS algorithm (Byrd et al. 1995).

4.2. Computational benchmarks

The computational expense of TOD simulations can vary significantly depending on the instrument characteristics and the desired level of detail. In general, the cost of TOD simulations for CMB experiments is largely independent of the final map resolution. Instead it is predominantly driven by the complexity of the

Table 1. Computational cost of simultaneous emulation on a batch of ten samples (each sample being made of a Q and a U polarization map) as a function of map resolution, N_{side} .

N_{side}	Memory [GB]	GPU hrs
16	0.419	0.12
32	1.512	0.25
64	6.151	0.46
128	25.93	1.11
256	95.96	2.17
512	376.7	4.5
1024	1478	9.3
2048	5804	19.3

Notes. The table shows the peak memory usage in GB and speed in GPU hours for each resolution level, obtained using double precision floats on an NVIDIA A100 64 GB GPU. Values of up to $N_{\text{side}} = 128$ were measured, while the ones at a higher resolution were obtained extrapolating the trend of the measured points (see Section 4.2 for details).

systematics effects and noise included in the timeline and by the assumptions made in the calibration and map-making steps, such as the correlation length of the noise assumed in the map-making (Wright et al. 1996; Tegmark 1997a). In many past efforts, massive Monte Carlo simulation campaigns have consumed $O(100)$ million CPU hrs to produce only $O(10^3)$ high-resolution ($N_{\text{side}} = 2048$) simulations (see e.g. Planck Collaboration XII 2016), amounting roughly to $O(10^5)$ CPU hrs per map.

The augmentation approach advocated here bypasses additional TOD generation and the subsequent map-making steps¹¹ by directly emulating at map-level, once the generative model has been built on a small ensemble of simulations. In contrast to TOD simulations, the emulator offers the possibility of directly generating low-resolution maps where appropriate, resulting in significant saving in computational resources. Its computational cost and memory requirements scale linearly with the number of orientations, R , and number of wavelet scales, J , employed in the analysis. Table 1 summarizes the computational and memory costs of the emulator at several N_{side} values on a single NVIDIA A100 64 GB GPU, assuming that the gradient descent is performed simultaneously on a batch of ten maps for 1000 steps on the four losses described in Section 3.4.2, using double precision floats¹². Values in this table for resolutions up to $N_{\text{side}} = 128$ are measured, and then extrapolated to higher resolution by fitting a power law of the kind $x(N_{\text{side}}) = A * (N_{\text{side}})^a$ (where x is either the cost in GPU hours or the memory in GB, A , and a are constants estimated during the fit) to the measured data. The estimated execution time (in GPU hrs) to generate ten new emulated maps scales roughly linearly with N_{side} and is significantly smaller than the respective cost of a TOD simulation in CPU hrs, even at a very high resolution (e.g., compare 1.93 GPU hrs per map at $N_{\text{side}} = 2048$ versus $O(10^5)$ CPU hrs for TOD). While emulation of high-resolution maps (e.g., $N_{\text{side}} = 2048$) might still incur a significant memory cost, most practical applications – especially the ones focusing on B -mode polarization detection in full-sky

¹¹ We note that also the computational cost of map-making might be relevant if not comparable with the one of TOD production. The approach proposed in this paper can in principle avoid also further costs associated to this step.

¹² Using single precision would imply approximately a reduction of a factor of two in GPU hrs and memory requirements. Also, using short floats would reduce the requirements shown in Table 1 by a factor of four and might be a viable alternative in some cases.

⁷ <https://github.com/jmdelouis/HealpixML>

⁸ Specifically, implementing the WST (see Section 1).

⁹ <https://github.com/pcampeti/CMBS CAT/>

¹⁰ A jupyter notebook demo for CMBS CAT is available here.

experiments – do not require such a high resolution. The computational cost is dominated by the wavelet convolution at each iteration during gradient descent, which requires a large graph in memory to be computed efficiently. For high-resolution cases, a multi-scale approach (e.g. Marchand et al. 2023), in which large scales are computed first and small scales are computed locally, should be considered instead, significantly reducing the memory footprint below what is shown in Table 1 for large N_{side} values. We leave further exploration of this point to future work.

Finally, our current implementation leverages the MPI-parallelized design of HealpixML to efficiently scale on multiple CPUs or GPUs. Although the primary computational bottleneck remains the wavelet convolution in real space on the sphere (which is considerably faster when performed instead on a small flat-sky patch) at each iteration during gradient descent, we expect that continued development and optimization – guided by the requirements on augmentation on noise and systematics templates from collaborations such as *LiteBIRD*, *SO*, and *CMB-S4* – will further enhance the efficiency of our simulation framework.

5. Validation

In this section we present the visual and statistical validation of the generative model. First, we define the various datasets used in the validation strategy. Next, we introduce the statistical tests employed to compare the model outputs with the target distribution. Finally, we discuss the results of the validation campaign.

5.1. Datasets definition

The validation framework relies on seven different datasets, each playing a distinct role in testing different aspects of the generative model. We define them in the following.

Reference dataset. We began with a set of 10 000 end-to-end TOD simulations (as is described in Section 2), from which we extracted 10 000 corresponding Q and U maps. We call this the original dataset in the following. We used the original dataset exclusively to estimate the full QU pixel-to-pixel covariance and pixel-to-pixel correlation matrices, hereafter referred to as the reference covariance and correlation matrices. These matrices serve as the ground truth for the covariance (and correlation) structure we aim to reconstruct. We then generated a reference dataset of 10 000 maps as random Gaussian samples from this reference covariance. Given that these realizations are sampled from the reference covariance, we can ignore the sample variance introduced by estimating an empirical 6144×6144 covariance matrix from only 10 000 simulations and treat the reference covariance as the target covariance that the emulator should ideally reproduce (i.e., the ground-truth). This choice is further justified by the fact that the reference covariance matrix appears to encode quite well the pixel-to-pixel spatial information in the original dataset: this can be understood considering that we simulated the original dataset as an additive random Gaussian fluctuation in the TOD. We therefore do not expect significant differences in our results when emulating the original dataset or the reference dataset. We further detail the reasons behind this choice in Appendix D.

The reference dataset provides a basis for comparing various statistical properties and ensuring that the emulator captures the true underlying distribution described by the reference covariance. We emphasize that generating the reference dataset

as Gaussian realizations from the pixel-to-pixel reference covariance matrix results in an inhomogeneous Gaussian field that cannot be fully characterized by its power spectrum alone, since the latter lacks any spatial information related to the specific information at each pixel. The additional “beyond power spectrum” information encoded in the SC thus plays a crucial role in constraining the realizations.

Input datasets. To mimic realistic scenarios in which only a limited number of simulations are available, we generated two input datasets by drawing random Gaussian realizations from the reference covariance:

1. A small input dataset containing $N_{\text{input}} = 10$ realizations.
2. A moderate size input dataset containing $N_{\text{input}} = 100$ realizations.

As for the reference dataset, since these realizations were also sampled from the reference covariance, the latter represents the ground-truth covariance that the emulator should aim to reproduce. For each input dataset, we computed the empirical pixel covariance (and correlation) matrices (input covariance (correlation) matrix). Since the number of simulations used to estimate them is lower than the number of their degrees of freedom, the input covariances will be plagued by high sample variance and will be rank-deficient (i.e. under-determined), leading to instability under inversion (Hartlap et al. 2007).

Emulated datasets. We applied the generative model to each of the two input datasets, following the average-target strategy for the $N_{\text{input}} = 10$ set and the single-target one for $N_{\text{input}} = 100$ (see Section 3.4.1) and augmenting each to create two corresponding emulated datasets of 10 000 maps each. From these maps, we computed the pixel covariance and correlation matrices – hereafter referred to as the emulated covariance (correlation) matrices – to evaluate how accurately the emulator reconstructs the target covariance.

Gaussian datasets. For additional comparison, we also constructed two Gaussian datasets using a naive augmentation strategy. Specifically, from the isotropic angular power spectra of each map in the 10-map and 100-map input datasets¹³, we generated 1000 and 100 Gaussian random realizations, respectively, leading to two datasets of 10 000 maps each. The corresponding covariance matrix is denoted as the Gaussian covariance. In summary, the validation involves seven different datasets: a reference dataset, two input sets, two emulated sets, and the two Gaussian datasets.

5.2. Visual and statistical tests

We employed visual and statistical tests to verify that the generated maps are consistent with the true underlying distribution.

1. Visual inspection. We visually compared both the emulated map samples (Fig. 1) and emulated pixel covariances (Figs. 4 and 5) with the reference ones.
2. Angular power spectra. We compared the EE , BB , and EB power spectra of the emulated maps against the ones computed from the reference dataset (see Fig. 3). We expect the emulator to be able to reproduce both the average power spectrum and the correct variance-per-multipole.
3. Pixel covariance matrices. We visually and quantitatively compared the empirical pixel correlation matrices of the

¹³ We find completely similar behavior of the Gaussian correlation matrices as the ones shown in Figs. 4 and 5 in Section 5 when generating maps from the average power spectrum of the input dataset instead.

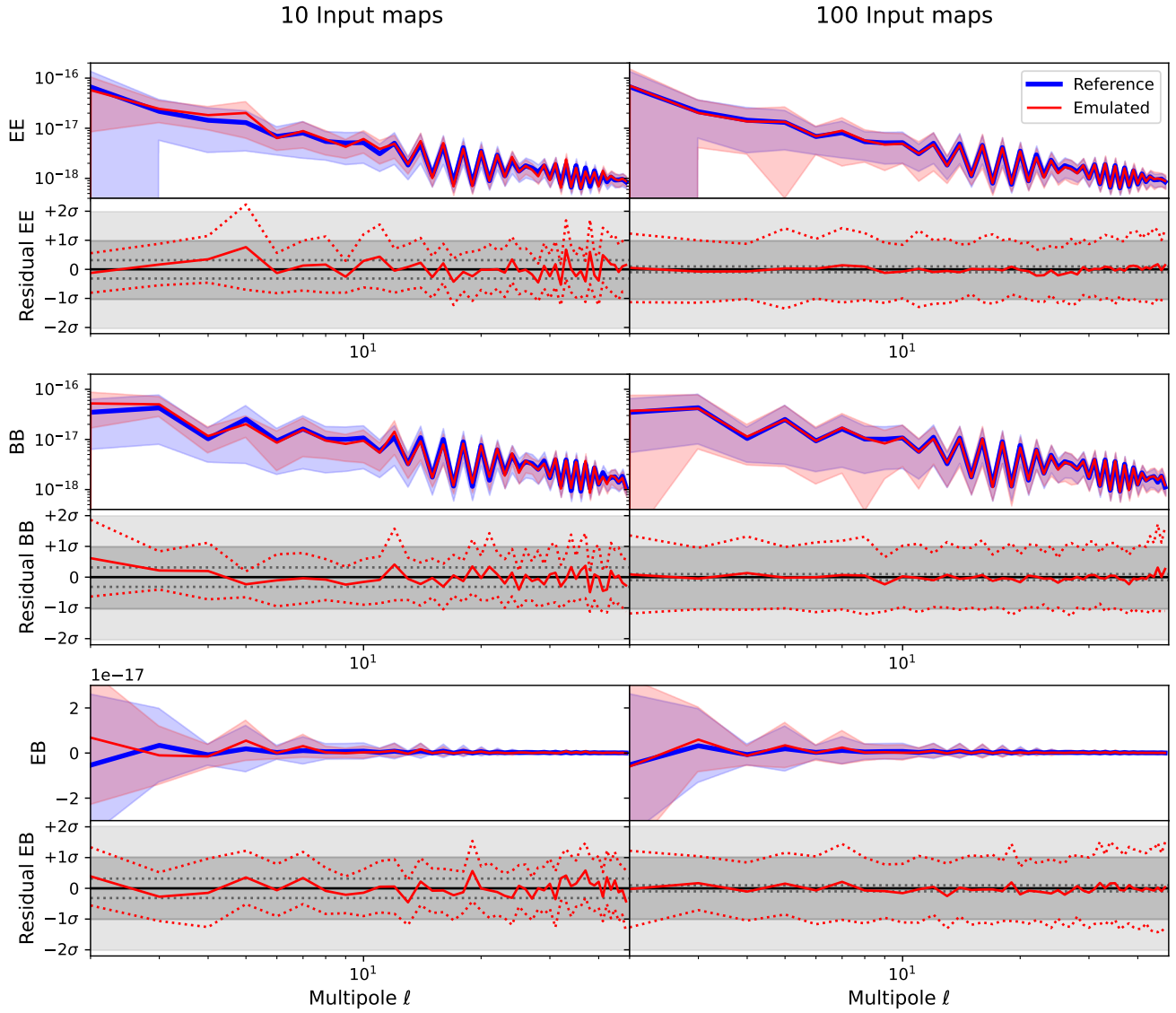


Fig. 3. Comparison of the EE , BB , and EB angular power spectra for the emulated maps (red) versus the reference dataset (blue), shown for the $N_{\text{input}} = 10$ (left) and $N_{\text{input}} = 100$ (right) cases. In each row, the top panel displays the average power spectra with 1σ shading, while the bottom panel shows the residual, $\Delta_\ell = (\langle C_\ell^{\text{emu}} \rangle - \langle C_\ell^{\text{ref}} \rangle) / \sigma_\ell^{\text{ref}}$ (solid red) along with $\Delta_\ell \pm \sigma_\ell^{\text{ref}}$ (dotted red), where σ_ℓ^{ref} is the standard deviation from the reference set and σ_ℓ^{emu} is the emulated dataset's one. In both the 10-map and 100-map cases, the mean spectra and variance per multipole closely match those of the reference set, and the residuals lie well within the 1σ band. The bottom panels also show in dotted gray the standard error on the mean for a sample of size N_{input} , that is $\pm \sigma_{\text{ref}} / \sqrt{N_{\text{input}}}$, with $N_{\text{input}} = 10$ (100) in the left (right) plots. Notably, the scatter of the mean residual is roughly within this band in both cases, indicating the success of the augmentation process. See Section 5.3 for further details.

emulated datasets with the reference one (Figs. 4 and 5). In particular, we examined the ratio of the eigenvalues of the covariance matrices (Fig. 6) and performed a χ^2 test (see point 4 below) to assess the fidelity of the off-diagonal structure.

4. Reduced χ^2 statistics. We computed the reduced χ^2 statistic for 1000 maps in the reference set using each covariance matrix:

$$\chi^2 = \frac{\mathbf{m}^T \mathbf{C}^{-1} \mathbf{m}}{d}, \quad (8)$$

where $\mathbf{m}$ represents a map from the reference dataset, $\mathbf{C}$ is one of the covariance matrices (reference, input, emulated, or Gaussian), and d is the number of degrees of freedom. In practice, we performed a principal component analysis (PCA) on $\mathbf{C}$ and retained the largest 150 eigenvalues of the

reference, emulated, and Gaussian covariances (accounting for approximately 90% of the total variance) and the largest 10 or 100 eigenvalues of the input covariance (depending whether $N_{\text{input}} = 10$ or 100), when computing χ^2 . The histogram of these reduced χ^2 values (see Fig. 6) provides another quantitative measure of statistical consistency.

We applied all of these tests to both the small (ten maps) and moderate (100 maps) input datasets. Additional validation metrics, such as the SC coefficients, Minkowski functionals, and probability distribution functions (PDFs) of the maps, are presented in Appendix C.

5.3. Results

The results shown in this section were obtained using the average-target strategy for the $N_{\text{input}} = 10$ case and the

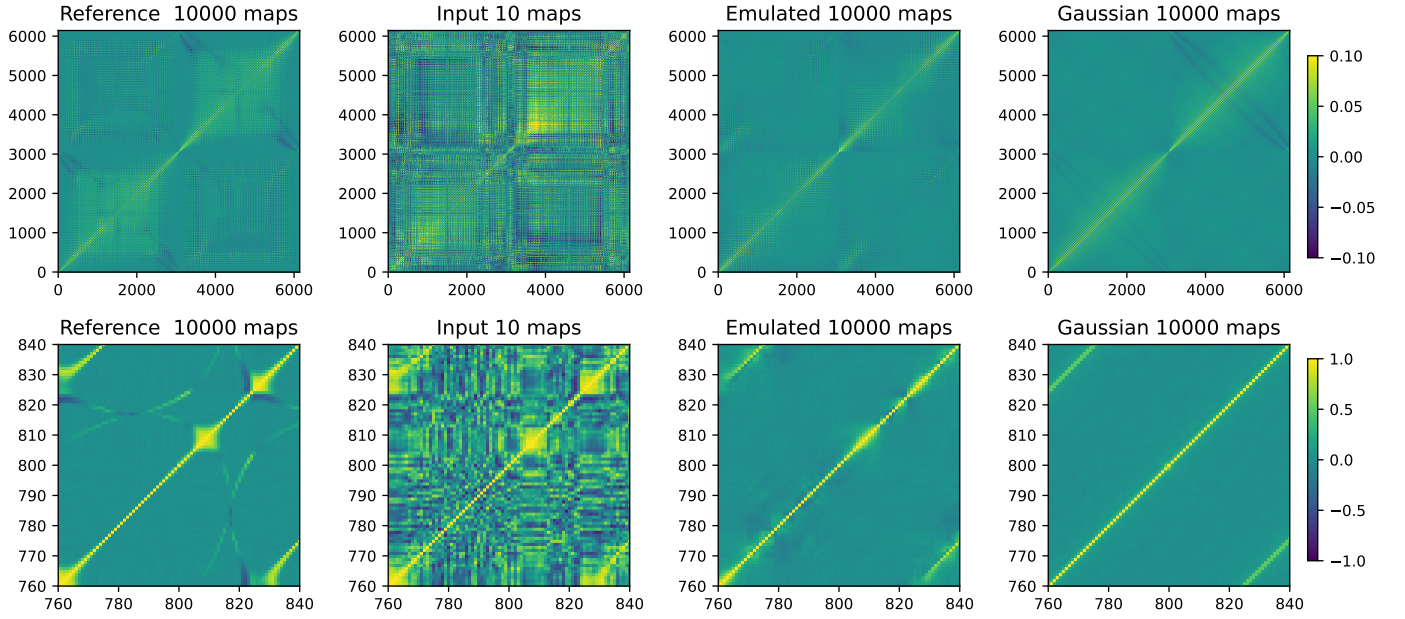


Fig. 4. Upper panel: empirical pixel correlation matrices estimated from the reference dataset, the input dataset, the emulated dataset, and the Gaussian dataset in the $N_{\text{input}} = 10$ case. The color scale runs between -0.1 and 0.1 to highlight the structures and allow for a visual comparison. See Section 5 for details. Bottom panels: zoom-in between pixel 760–840 of the corresponding covariance matrix in the upper panel. Note that here the color scale runs between -1 and 1 . See Section 5.3 for further details.

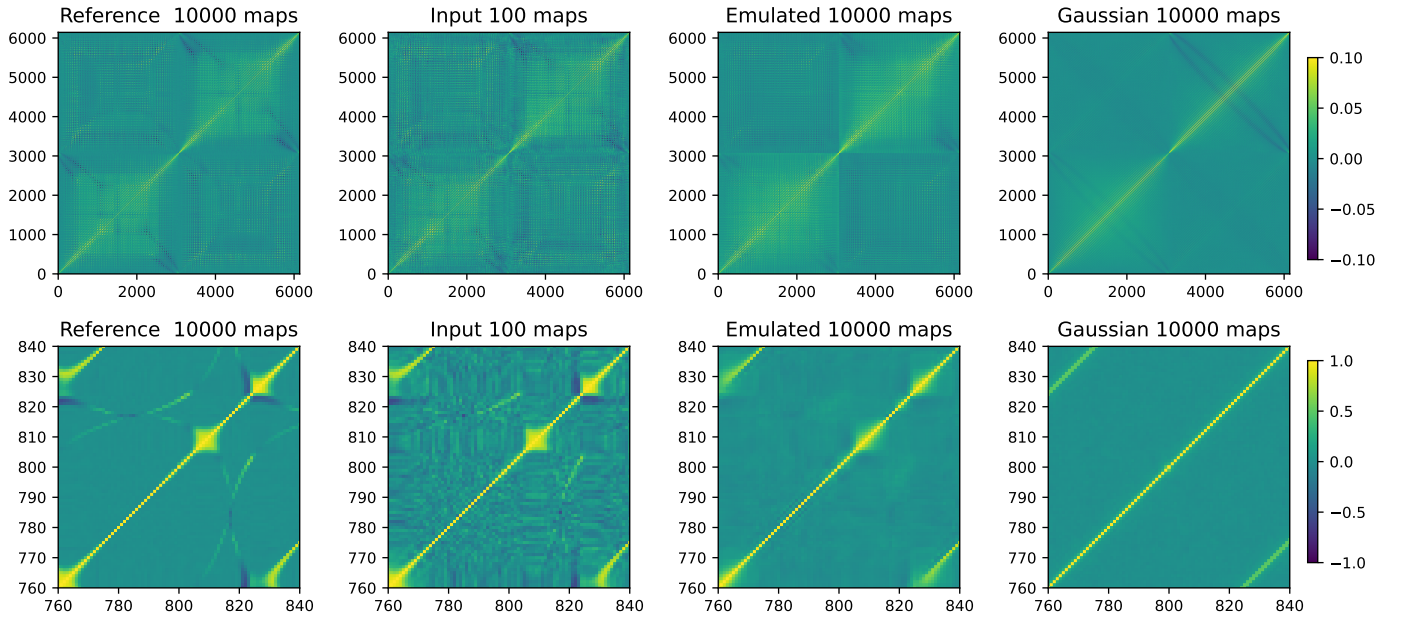


Fig. 5. Same as Figure 4 but for the $N_{\text{input}} = 100$ case. See Section 5.3 for further details.

single-target strategy for the $N_{\text{input}} = 100$ one (see Section 3.4.1). We find that the optimal choice of target for gradient descent depends on the size of the input dataset. When only a few tens of simulations (or fewer) are available, using the average-target approach (Eq. (4)) yields more stable results than the single-target strategy. This is because, in our case, a small sample is sufficient to capture somewhat accurately the “true” average of the SC coefficients across the reference dataset, leading to visually improved empirical covariance matrices and an enhanced agreement in both covariance eigenvalues and χ^2 values with the reference. Also for datasets comprising hundreds of simulations, using the average-target approach allows for improvements

across all metrics described in Section 5.2 when compared to the same quantities evaluated without any augmentation. However, in this case the single-target method (Eq. (5)) allows for further improvements over the average-target results. This is because the single-target strategy more effectively transfers the full statistical distribution encoded in the input dataset to the emulated outputs, capturing true structures in the covariance matrix beyond the capabilities of the average-target.

The results of the validation are organized into four main parts, following Section 5.2: visual inspection of the maps, power spectrum consistency, pixel-to-pixel covariance matrix reconstruction, and reduced χ^2 -test in pixel space.

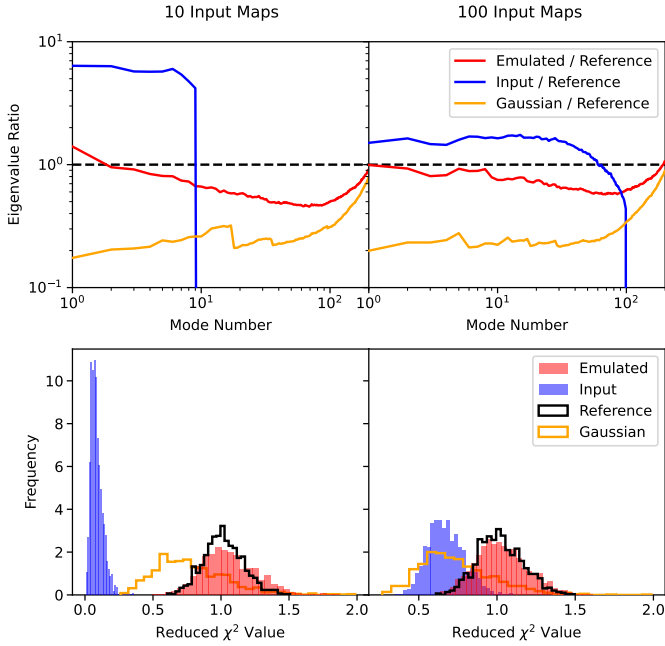


Fig. 6. Top panel: ratio of eigenvalues of the input, emulated, and Gaussian pixel covariance matrices in Fig. 4 to the reference covariance’s ones. Bottom panel: histograms of chi-square values computed using 1000 maps from the reference dataset and the four different covariance matrices. See Section 5.3 for details.

Visual inspection. Figure 1 shows a side-by-side comparison of three input maps and three corresponding emulated samples synthesized from each of the inputs. Specifically, we show in this figure the emulated maps obtained from the moderate-sized input dataset. For each row, the leftmost column displays the input (target) map from this dataset, while the other columns present distinct emulated samples obtained from the specific target map. We get a very similar visual comparison when emulating instead from the small (ten-maps) input set, which we do not show here. Notably, the emulated maps capture key features of the target, such as the scanning strategy stripes and the ecliptic pole patterns, at the right positions, while also displaying clear differences, similar to the ones between the different input maps. This suggests that the emulator successfully balances predictiveness (accurate feature reconstruction) with representativeness (maintaining sample variability).

Angular power spectra. In Fig. 3, we compare the EE , BB , and EB angular power spectra of the emulated maps to the ones from the reference dataset. To quantify their difference, we defined the residual

$$\Delta_\ell = \frac{\langle C_\ell^{\text{emu}} \rangle - \langle C_\ell^{\text{ref}} \rangle}{\sigma_\ell^{\text{ref}}},$$

where $\langle C_\ell^{\text{emu}} \rangle$ and $\langle C_\ell^{\text{ref}} \rangle$ are the average power spectra in the emulated and reference sets, respectively, and σ_ℓ^{ref} is the standard deviation measured in the reference set. As is shown in the bottom panels (solid red lines), Δ_ℓ remains well within the 1σ band across the multipoles. To compare the variance-per-multipole in the two datasets, we also plot (as dotted red lines)

$$\Delta_\ell^\pm = \frac{\langle C_\ell^{\text{emu}} \rangle - \langle C_\ell^{\text{ref}} \rangle \pm \sigma_\ell^{\text{emu}}}{\sigma_\ell^{\text{ref}}},$$

where σ_ℓ^{emu} is the standard deviation in the emulated dataset. Both the small and moderate input cases agree very well with the reference set in terms of mean spectra and variance. Notably, even when trained on only ten input maps, the emulator still reproduces the mean and variance of the angular power spectrum with good fidelity. We also plot in the bottom panels of Fig. 3 two dotted gray lines indicating the standard error on the mean for a sample of size N_{input} , namely $\pm \sigma_{\text{ref}} / \sqrt{N_{\text{input}}}$, with $N_{\text{input}} = 10$ in the left plots and $N_{\text{input}} = 100$ in the right plots. Notably, the scatter on the mean residual is roughly within this band in both cases, indicating that the augmentation process is approximately at least as good as having N_{input} new direct draws from the reference dataset: the emulated dataset reproduces the mean power spectra as closely as one would expect if one had roughly N_{input} truly independent draws from the true distribution, without adding any significant extra bias beyond that expected from the small input dataset.

Pixel covariance matrices. Figures 4 and 5 show the empirical pixel correlation matrices for the reference, input, and emulated datasets¹⁴. In the ten-map case, the input correlation is visibly noisy, reflecting the small sample size. However, the emulated correlation matrix – computed from 10 000 generated maps – exhibits a much cleaner structure, while retaining some of the main features of the reference correlation matrix. With 100 maps, the input correlation matrix suffers less from sample noise, but the emulated one continues to capture both prominent and more subtle features seen in the reference significantly better than the input one. We note that some of the more subtle “filamentary” structures in reference covariance in the bottom panel of Fig. 5 appear slightly smoothed over more pixels in the emulated one. This is connected to the number of wavelet orientations, R , and the size of the wavelet convolution kernel (i.e., 3×3) we chose (see Section 3.2). While this specific choice already gives satisfactory results in all key metrics considered, we note that it is sufficient to increase both of these parameters to recover these subtle off-diagonal features more sharply.

Eigenvalues and reduced χ^2 statistics. To quantify these observations, we compare in Fig. 6 the eigenvalues of the different covariance matrices and the reduced χ^2 values for 1000 maps in the reference set using each covariance matrix (Eq. (8)). These comparisons again indicate that the emulated maps help to more accurately reproduce the true underlying distribution compared to the naive estimation from 10 or 100 input maps. Because the generative model can produce new (approximate) independent realizations of the underlying stochastic process, augmenting the input dataset reduces the sample variance and, consequently, the bias it introduces in the inverse covariance matrices (Hartlap et al. 2007). Therefore, the eigenvalues and the reduced χ^2 distribution derived from the emulated covariance matrix match more closely the ones from the reference covariance, surpassing the performance of the covariance naively estimated only from input simulations, in both the $N_{\text{input}} = 10$ and the $N_{\text{input}} = 100$ cases. In particular, the top panels in Fig. 6 show that the generative model-driven augmentation helps recover the first ~ 150 eigenvalues significantly better than the naive estimation, while the bottom panels illustrate that the reduced χ^2 values from the emulated covariance cluster around 1, and have a distribution consistent with the reference. In contrast, the reduced χ^2 values from the input covariance cluster are consistently off

¹⁴ We show correlation matrices obtained from maps in HEALPix RING ordering.

from 1, indicating a very poor estimation of uncertainties in the reference set.

In addition to the emulation-based approach, we also considered a naive augmentation strategy – labeled “Gaussian” in Figs. 4–6 – which consists of generating additional maps by sampling new Gaussian random realizations from the isotropic power spectrum of each input map¹⁵. Although this approach can produce visually “cleaner” off-diagonal structures in the covariances (see the rightmost panels of Figs. 4 and 5), it fails to replicate even the brighter, more pronounced features observed in the reference correlation matrix. Comparisons of both eigenvalues and the reduced χ^2 histograms (Fig. 6) confirm that this naive augmentation method falls short of capturing the full covariance structure and performs significantly worse than the proposed emulation-based strategy.

6. Conclusions

In this work, we have introduced a novel emulator for CMB systematics datasets based on SC statistics. Unlike neural-network models developed for similar purposes (e.g., diffusion models, Sohl-Dickstein et al. 2015), this approach is fully interpretable and requires only a small number of high-fidelity simulations to generate new accurate and statistically consistent samples. We demonstrate that a small ensemble can be effectively augmented with such an emulator to achieve a robust statistical performance in several key tasks of cosmological analyses. We deployed a number of tests in order to demonstrate this statistical robustness, among them the mean and variance of power spectrum estimates and the reduction in sample variance and its associated bias in inverse pixel covariance matrices.

The emulator was validated using realistic, highly inhomogeneous map simulations of CMB systematics, which are especially challenging to emulate. We tested the emulator on both small (ten input maps) and moderate-sized (100 input maps) datasets. As was expected, increasing the number of input simulations improves the quality of emulation. Nonetheless, even when trained on just ten simulations, the emulator produces remarkable results in both visual and statistical tests.

Although generative models based on the SC statistics naturally target homogeneous data, we nonetheless achieve successful emulation in this more complex context, characterized by strongly inhomogeneous statistics. Consequently, we are confident that the SC approach – especially when complemented by the orientation-informed wavelets that we introduced here – can be applied to a broad range of observational scenarios featuring similar large-scale instrumental structures.

The framework can also be adapted to multiplicative or signal-dependent systematics, beyond the additive noise example studied here. Instead of performing data augmentation by sampling independent realizations of the noise, we should instead recast the task as an inverse problem, describing the multiplicative systematic as the action of a stochastic forward model that maps a clean signal to a contaminated one. This can be achieved, for instance, by estimating the joint and cross scattering statistics of the signal and the contaminant, and then sampling new noise realizations conditioned on a given signal map. Multichannel generative models with cross ST statistics have already been introduced in Régalo-Saint Blancard et al. (2023). Also, Delouis et al. (2022) show an example of map sampling whose cross-statistics with another signal is constrained

(i.e. Q/U polarized dust emission at 353 GHz conditioned with a total intensity dust emission map).

The emulator can operate entirely on GPUs, which in this case are significantly more energy-efficient than CPUs, offering substantial practical benefits. For example, augmenting a dataset of 100 low-resolution Q and U maps (with $N_{\text{side}} = 16$) by a factor of 100 requires roughly 120 GPU-hours on a single NVIDIA A100 64 GB GPU – a negligible cost compared to the millions of CPU-hours typically needed for full end-to-end simulations in full-sky experiments. This might enable analyses on both data and simulations that were previously unfeasible due to a lack of computational resources. Moreover, the energy efficiency of GPUs further suggests that our method could significantly reduce the carbon footprint of extensive simulation campaigns.

Looking forward, we plan to also make our method efficient for higher-resolution maps and benchmark its performance in multi-GPU environments, which are necessary to overcome current GPU memory limitations during simultaneous gradient descent. Given the strong scaling of the framework to multiple GPUs, there is a clear road-map ahead for its development and testing. On the algorithmic front, incorporating advanced techniques such as meta-optimization may further enhance the stability and accuracy of the loss minimization process.

Moreover, our emulator opens the door to a range of applications beyond augmentation, including simulation-based inference for cosmological parameters and the denoising of complex instrumental systematics. Simulation-based inference might indeed prove necessary to fully exploit future cosmological datasets, and avoid the hard (or sometimes even impossible) task of characterizing complex systematics effects at the likelihood level (Wolz et al. 2023). Also quadratic maximum-likelihood (QML) power spectrum estimators and pixel-based likelihoods (Tegmark 1997b) might benefit from our approach: augmenting the datasets used to compute the empirical noise pixel covariance matrices needed by both methods might help to achieve optimal errors bars, especially in scenarios in which only a few hundred simulations are available. As a concrete example, we plan to use our method to mitigate ADC nonlinearities (Planck Collaboration Int. XLVI 2016) in Planck polarization data (see e.g. Wolz et al. 2023, for a simulation-based approach), aiming for more accurate estimates of the optical depth of reionization and tensor-to-scalar ratio.

In summary, our approach advocates for a paradigm shift in simulation strategies: high-fidelity emulation does not require an extensive number of simulations, but rather a few carefully crafted ones that accurately represent the experimental noise model and sky. Rather than producing vast numbers of lower- or intermediate-quality simulations, future pipelines for cosmological surveys – such as Euclid, *LiteBIRD*, SO, CMB-S4, and Rubin-LSST – should focus on generating a few high-precision simulations. These can then be efficiently augmented using the framework presented here.

Data availability

In the spirit of reproducibility and accessibility, we provide to the community both a general framework for computing the SC statistics HealpixML at <https://github.com/jmdelouis/HealpixML> and the emulator CMBSCAT at <https://github.com/pcampeti/CMBSCAT/> we developed.

Acknowledgements. We thank: V. Amitrano, G. Galloni and W.L. Kimmy Wu for advice and fruitful discussions; E. Calore and A. Miola for help in setting up the computational environment. We acknowledge the use of computing

¹⁵ We find completely similar results when generating maps from the average power spectrum of the input dataset instead.

facilities provided by the INFN theory group (I.S. InDark) at CINECA and the COKA machine at INFN Sezione di Ferrara. This work has been supported by the Fondazione ICSC, Spoke 3 Astrophysics and Cosmos Observations. National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) Project ID CN_00000013 “Italian Research Center on High-Performance Computing, Big Data and Quantum Computing” funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di “campioni nazionali di R&S (M4C2-19)” - Next Generation EU (NGEU). M.G., M.L. and L.P. acknowledge the financial support from the INFN InDark initiative and from the COSMOS network through the ASI (Italian Space Agency) Grants 2016-24-H.0 and 2016-24-H.1-2018. M.G. and M.L. are funded by the European Union (ERC, RELICS, project number 101116027). M.G. acknowledges support from the PRIN (Progetti di ricerca di Rilevante Interesse Nazionale) number 2022WJ9J33.

References

- Abazajian, K. N., Adshead, P., Ahmed, Z., et al. 2016 arXiv e-prints [arXiv:1610.02743]
- Acosta, M. C., Palomas, S., Paronuzzi Ticco, S. V., et al. 2024, *GMD*, **17**, 3081
- Allys, E., Levrier, F., Zhang, S., et al. 2019, *A&A*, **629**, A115
- Allys, E., Marchand, T., Cardoso, J. F., et al. 2020, *Phys. Rev. D*, **102**, 103506
- Auclair, C., Allys, E., Boulanger, F., et al. 2024, *A&A*, **681**, A1
- Beck, D., Cukierman, A., & Wu, W. L. K. 2022, *MNRAS*, **515**, 229
- BICEP collaboration 2021, *Phys. Rev. Lett.*, **127**, 151301
- Blancard, R.-S., et al. 2024, *Phys. Rev. D*, **109**, 083535
- Bruna, J., & Mallat, S. 2013, *PAMI*, **35**, 1872
- Bruna, J., & Mallat, S. 2019, *Math. Stat. Learn.*, **1**, 257
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. 1995, *SISC*, **16**, 1190
- Carones, A., CarrónDuque, J., Marinucci, D., Migliaccio, M., & Vittorio, N. 2024, *MNRAS*, **527**, 756
- Carrón Duque, J., Carones, A., Marinucci, D., Migliaccio, M., & Vittorio, N. 2024, *JCAP*, **01**, 039
- Chardin, J., Uhlrich, G., Aubert, D., et al. 2019, *MNRAS*, **490**, 1055
- Cheng, S., & Ménard, B. 2021, arXiv e-prints [arXiv:2112.01288]
- Cheng, S., Ting, Y.-S., Ménard, B., & Bruna, J. 2020, *MNRAS*, **499**, 5902
- Cheng, S., Morel, R., Allys, E., Ménard, B., & Mallat, S. 2024, *PNAS Nexus*, **3**, 103
- Cranmer, K., Brehmer, J., & Louppe, G. 2020, *PNAS*, **117**, 30055
- Delouis, J. M., Pagano, L., Mottet, S., Puget, J. L., & Vibert, L. 2019, *A&A*, **629**, A38
- Delouis, J. M., Allys, E., Gauvrit, E., & Boulanger, F. 2022, *A&A*, **668**, A122
- Eickenberg, M., Allys, E., Dizgah, A. M., et al. 2022, arXiv e-prints [arXiv:2204.07646]
- Gatti, M., Jeffrey, N., Whiteway, L., et al. 2024, *Phys. Rev. D*, **109**, 063534
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. 2014, arXiv e-prints [arXiv:1406.2661]
- Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, *ApJ*, **622**, 759
- Häggbom, M., Karlsmark, M., & Andén, J. 2024, arXiv e-prints [arXiv:2403.08362]
- Hartlap, J., Simon, P., & Schneider, P. 2007, *A&A*, **464**, 399
- Jeffrey, N., Boulanger, F., Wandelt, B. D., et al. 2022, *MNRAS*, **510**, L1
- Kingma, D. P., & Welling, M. 2013, arXiv e-prints [arXiv:1312.6114]
- Ledoit, O., & Wolf, M. 2003, Honey, I shrunk the sample covariance matrix, Economics Working Papers 691, Department of Economics and Business, Universitat Pompeu Fabra
- LiteBIRD Collaboration 2023, *PTEP*, **2023**, 042F01
- Mallat, S. 2012, *Comm. Pure Appl. Math.*, **65**, 1331
- Marchand, T., Ozawa, M., Biroli, G., & Mallat, S. 2023, *Phys. Rev. X*, **13**, 041038
- Morel, R., Rochette, G., Leonarduzzi, R., Bouchaud, J.-P., & Mallat, S. 2022, arXiv e-prints [arXiv:2204.10177]
- Mousset, L., Allys, E., Price, M. A., et al. 2024, *A&A*, **691**, A269
- Ng, A., & Jordan, M. 2001, in *Advances in Neural Information Processing Systems*, eds. T. Dietterich, S. Becker, & Z. Ghahramani (Cambridge: MIT Press), 14
- Planck Collaboration XII. 2016, *A&A*, **594**, A12
- Planck Collaboration VI. 2020, *A&A*, **641**, A6
- Planck Collaboration LVII. 2020, *A&A*, **643**, A42
- Planck Collaboration Int. XLVI. 2016, *A&A*, **596**, A107
- Price, M. A., Mars, M., Docherty, M. M., et al. 2023, arXiv e-prints [arXiv:2307.04798]
- Régaldo-Saint Blancard, B., Levrier, F., Allys, E., Bellomi, E., & Boulanger, F. 2020, *A&A*, **642**, A217
- Régaldo-Saint Blancard, B., Allys, E., Boulanger, F., Levrier, F., & Jeffrey, N. 2021, *A&A*, **649**, L18
- Régaldo-Saint Blancard, B., Allys, E., Auclair, C., et al. 2023, *ApJ*, **943**, 9
- Rodriguez, A. C., Kacprzak, T., Lucchi, A., et al. 2018, *Comput. Astrophys. Cosmol.*, **5**, 4
- Sellentin, E., & Heavens, A. F. 2016, *MNRAS*, **456**, L132
- Simons Observatory 2019, *JCAP*, **02**, 056
- Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. 2015, arXiv e-prints [arXiv:1503.03585]
- Tauber, J. A., Mandolesi, N., Puget, J.-L., et al. 2010, *A&A*, **520**, A1
- Tegmark, M. 1997a, *Phys. Rev. D*, **56**, 4514
- Tegmark, M. 1997b, *ApJ*, **480**, L87
- Tristram, M., et al. 2022, *Phys. Rev. D*, **105**, 083524
- Valogiannis, G., & Dvorkin, C. 2022, *Phys. Rev. D*, **105**, 103534
- Valogiannis, G., Yuan, S., & Dvorkin, C. 2024, *Phys. Rev. D*, **109**, 103503
- Wolz, K., Krachmalnicoff, N., & Pagano, L. 2023, *A&A*, **676**, A30
- Wright, E. L., Hinshaw, G., & Bennett, C. L. 1996, *ApJ*, **458**, L53
- Zhang, S., & Mallat, S. 2019, arXiv e-prints [arXiv:1911.10017]

Appendix A: Orientation-informed wavelets

A.1. Spherical wavelet definition

The SC statistics used in this study is based on oriented wavelet convolutions on the sphere. The convolution uses the HealpixML software¹⁶ based on HEALPix (Górski et al. 2005) to describe the pixelized data on the sphere. The wavelet kernel is different for each pixel as the HEALPix pixel could have a different shape, in particular near the pole and at large scale. Thus, the original complex kernel defined by the equation

$$W_{j,\Theta} = (\cos(2\pi\alpha) + j \sin(2\pi\alpha)) e^{-(\hat{\theta}^2 + \hat{\phi}^2)} \quad (\text{A.1})$$

$$N_{side} = 2^j \quad (\text{A.2})$$

$$\hat{\theta} = N_{side} \left(\theta - \frac{\pi}{2} \right)$$

$$\hat{\phi} = N_{side} \phi$$

$$\alpha = \hat{\phi} \cos \Theta + \hat{\theta} \sin \Theta$$

is convolved with the data to compute the wavelet of the direction Θ at the scale j .

Although this representation is locally valid, consistent wavelet computation on the sphere requires defining the absolute orientation Θ and determining meaningful directions. Figure A.1 demonstrates how the wavelet direction is specified. It is important to understand that a consistent orientation definition across the entire sphere is unattainable. The chosen definition perfectly aligns with the symmetric definition for the real part, where orientation does not affect the weights. However, changes in orientation alter the sign of the imaginary part. This definition renders the southern hemisphere definition as the conjugate of the northern hemisphere.

The direction of rotation does not affect the SC statistics, as applying the norm eliminates the influence of this rotation. This is because the SC statistics is computed after normalizing, which conceals the sign of the imaginary component, and also because all calculations are performed at the pixel level, ensuring consistency in the definition.

A.2. Rotation matrix definition

The SC coefficients are determined by integrating the local wavelet norm across the entire domain. In instances where there is a preferred orientation in a specific region, typically observed in data retrieved from scanning operations, this preferred orientation influences the data statistics. Consequently, the homogeneity of the entire domain is not ensured, as the dominant orientation at any given scale can vary throughout the domain. To address this, the orientation with the largest amplitude at each pixel is computed at each scale, and this orientation is then referred to as the primary orientation. The SC statistics are then defined relatively to the scanning strategy. To prevent instability, the determination of the maximum is smoothed to avoid rapid changes in the preferred orientation at each pixel. This rotation is applied after calculating the wavelet, resulting in four coefficients per pixel corresponding to the four rotations, which are computed using a precomputed 4×4 matrix for each pixel.

Figures A.2 present the power spectra of the directive wavelet. Despite the compact size of the kernel used to characterize the wavelet, the resulting power spectra separation is superior to 10^{-2} . Figure A.3 illustrates the complete signal decomposition, displaying the transfer function for the cumulative wavelets.

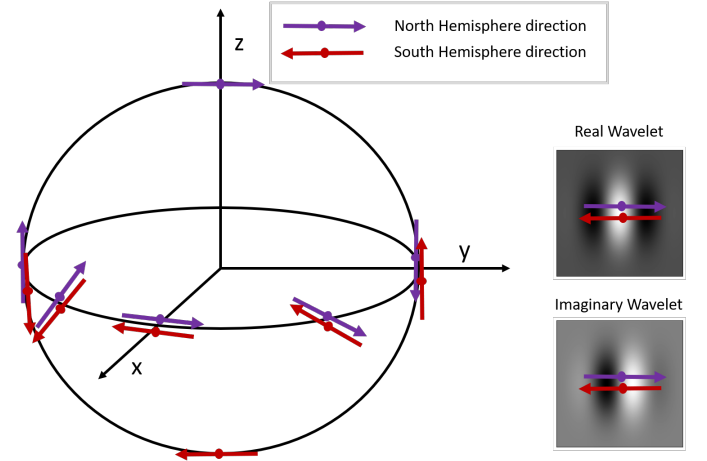


Fig. A.1. Explanation of the local rotational frame utilized to represent the wavelet coefficients. An inversion of direction occurs between the northern (depicted by the blue arrow) and southern (depicted by the red arrow) hemispheres. The effect of this directional inversion is shown in both the real part (top right panel) and the imaginary part (bottom right panel).

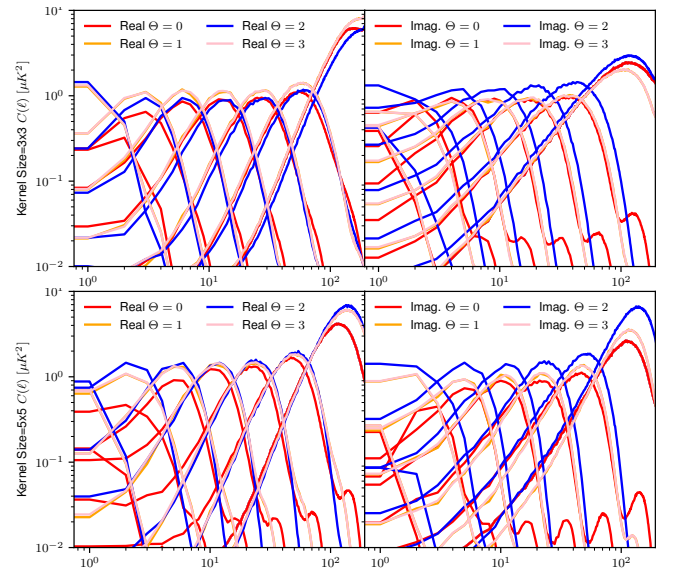


Fig. A.2. Power spectra from the wavelet transformation across the entire sky at $n_{side}=64$, spanning 7 scales. The upper two panels display results for a wavelet described by a 3×3 kernel size, whereas the lower two panels present identical results for a 5×5 kernel size. The panels on the left depict the power spectra of the real wavelet component, and those on the right illustrate the imaginary component.

This figure highlights the efficacy of the multiscale decomposition in covering nearly the entire scale domain. A 5×5 kernel faces greater challenges in accurately representing very small multipoles compared to a 3×3 kernel. This issue arises from the longer wavelet length, which requires more pixels for accurate depiction. Specifically, enlarging the mother-wavelet support from 3×3 to 5×5 pixels makes the coarsest dilated kernels span a sizeable fraction of the sky (~ 200 deg). On the discrete HEALPix grid this wide support is sampled and truncated less accurately (especially near the poles), so the cumulative transfer function drops below unity at $\ell \lesssim 10$. On the other hand, a 3×3 mother-wavelet avoids these issues while still covering the full angular range of interest through dyadic dilations.

¹⁶ <https://github.com/jmdelouis/HealpixML>

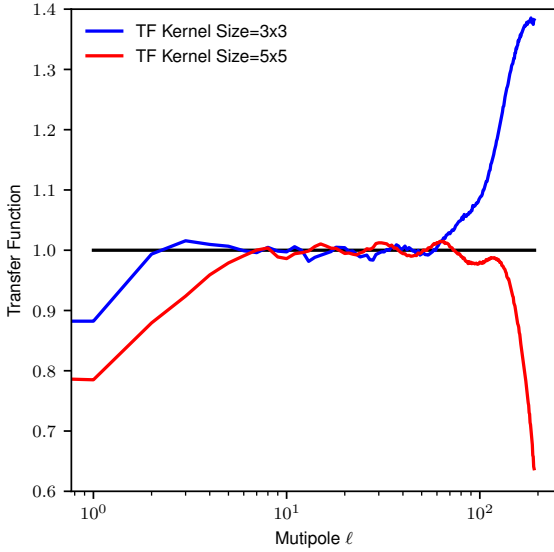


Fig. A.3. Examination of the cumulative transfer function from the wavelet transforms used in this research, derived from 100 samples of white noise. The blue (and red) curve illustrates the aggregate transfer function of the wavelet transforms in all directions and scales, corresponding to a kernel size of 3×3 (and 5×5 , respectively).

Appendix B: Cross scattering covariance statistics

We report here expressions for the cross SC statistics, extending the ones in Section 3.2 (Eqs. 1) to the cross-correlations between two different input fields I_1 and I_2 :

$$\begin{aligned} S_1^{\times, \lambda_1} &= \left\langle \left| (I_1 * \Psi_{\lambda_1})(I_2 * \Psi_{\lambda_1})^* \right| \right\rangle, \\ S_2^{\times, \lambda_1} &= \left\langle (I_1 * \Psi_{\lambda_1})(I_2 * \Psi_{\lambda_1})^* \right\rangle, \\ S_3^{\times, \lambda_1, \lambda_2} &= \text{Cov} \left[I_1 * \Psi_{\lambda_1}, |I_2 * \Psi_{\lambda_2}| * \Psi_{\lambda_1} \right], \\ S_{3p}^{\times, \lambda_1, \lambda_2} &= \text{Cov} \left[I_2 * \Psi_{\lambda_1}, |I_1 * \Psi_{\lambda_2}| * \Psi_{\lambda_1} \right], \\ S_4^{\times, \lambda_1, \lambda_2, \lambda_3} &= \text{Cov} \left[|I_1 * \Psi_{\lambda_3}| * \Psi_{\lambda_1}, |I_2 * \Psi_{\lambda_2}| * \Psi_{\lambda_1} \right], \end{aligned} \quad (\text{B.1})$$

where $*$ denotes convolution, covariances are defined as $\text{Cov}[X, Y] = \langle XY^* \rangle - \langle X \rangle \langle Y^* \rangle$ for two complex fields X and Y and $\langle \cdot \rangle$ indicates a spatial average.

Appendix C: SC coefficients, Minkowski functionals and PDFs

We complement the validation shown in Section 5 by plotting the SC coefficients (Fig. C.1), the Minkowski functionals (Fig. C.2) and the PDFs (probability density functions) in pixel space (Fig. C.3) of all the datasets considered here. As expected, we find remarkable agreement in all these test statistics.

Figure C.1 shows the SC coefficients for the reference (blue), emulated (red) and Gaussian (orange) datasets, for both the 10-maps (left panels) and 100-maps (right panels) input datasets. Each panel shows the average SC coefficient surrounded by a shaded area representing the standard deviation of the coefficients. The indices are ordered from the smallest to the largest convolved scales for each coefficient. To quantify differences between the SC coefficients of the reference and emulated

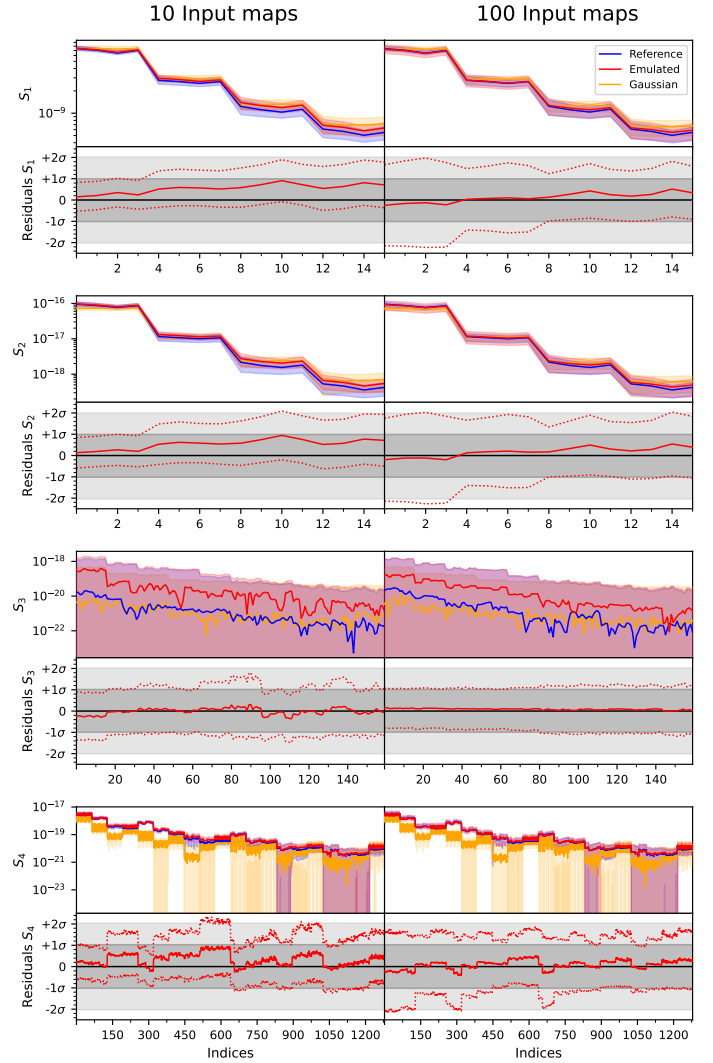


Fig. C.1. SC coefficients of the Q map for the reference (blue), emulated (red) and Gaussian (orange) datasets, for both the 10-maps (left panels) and 100-maps (right panels) input datasets. In each row, the top panel displays the average SC coefficient with $1\text{-}\sigma$ shading, while the bottom panel shows the residual, $\Delta S_X = (\langle S_X^{\text{emu}} \rangle - \langle S_X^{\text{ref}} \rangle) / \sigma_{S_X}^{\text{ref}}$ (solid red) along with $\Delta S_X^{\pm} = (\langle S_X^{\text{emu}} \rangle - \langle S_X^{\text{ref}} \rangle \pm \sigma_{S_X}^{\text{emu}}) / \sigma_{S_X}^{\text{ref}}$ (dotted red), where $\sigma_{S_X}^{\text{ref}}$ is the standard deviation from the reference set and $\sigma_{S_X}^{\text{emu}}$ is the emulated dataset's one.

datasets, we define their residuals as

$$\Delta S_X = \frac{\langle S_X^{\text{emu}} \rangle - \langle S_X^{\text{ref}} \rangle}{\sigma_{S_X}^{\text{ref}}},$$

where $X = 1, 2, 3, 4$ identifies the specific SC coefficient, $\langle S_X^{\text{emu}} \rangle$ and $\langle S_X^{\text{ref}} \rangle$ are the average coefficients in the emulated and reference sets, respectively, and $\sigma_{S_X}^{\text{ref}}$ is the standard deviation measured in the reference set. As shown in the bottom panels of Fig. C.1 (solid red lines), ΔS_X remains well within the 1σ band for all indices. To compare the variance in the two datasets, we also plot

$$\Delta S_X^{\pm} = \frac{\langle S_X^{\text{emu}} \rangle - \langle S_X^{\text{ref}} \rangle \pm \sigma_{S_X}^{\text{emu}}}{\sigma_{S_X}^{\text{ref}}},$$

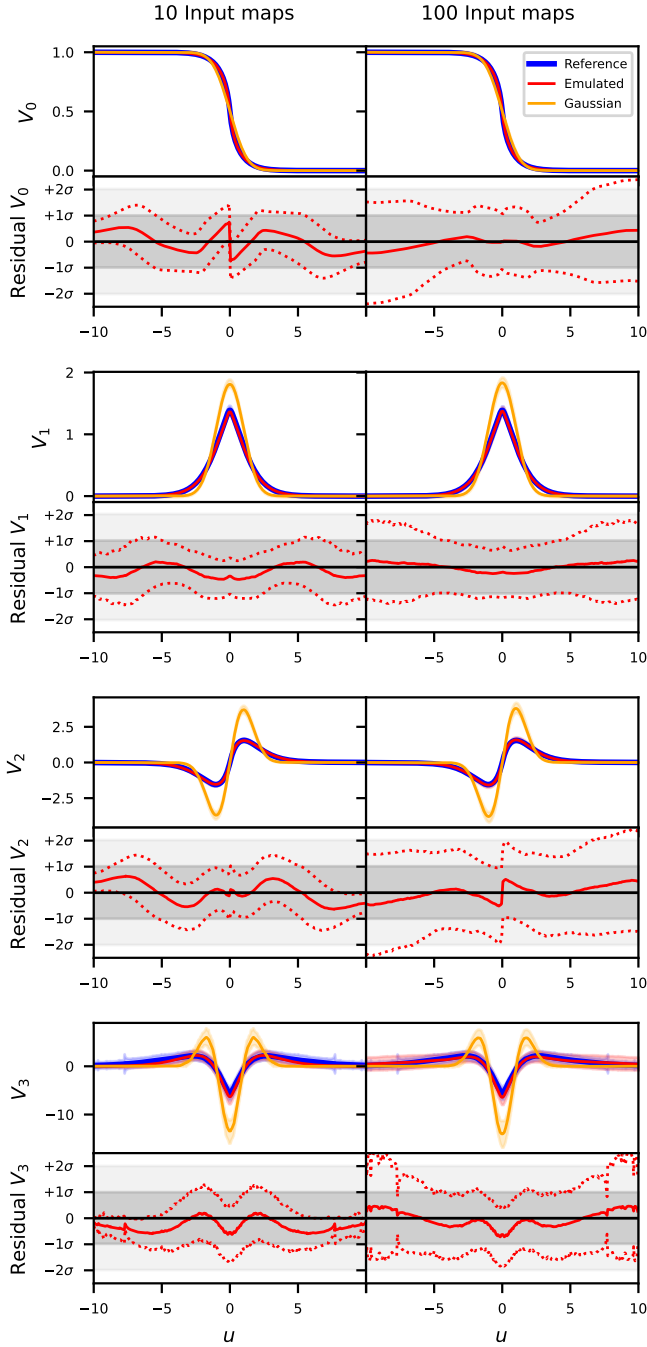


Fig. C.2. The four SO(3) Minkowski functionals for the reference (blue), emulated (red) and Gaussian (orange) datasets, for both the 10-maps (left panels) and 100-maps (right panels) input datasets. In each row, the top panel displays the average Minkowski functional with $1\text{-}\sigma$ shading, while the bottom panel shows the residual, $\Delta V_X = (\langle V_X^{\text{emu}} \rangle - \langle V_X^{\text{ref}} \rangle) / \sigma_{V_X}^{\text{ref}}$ (solid red) along with $\Delta V_X^{\pm} = (\langle V_X^{\text{emu}} \rangle - \langle V_X^{\text{ref}} \rangle \pm \sigma_{V_X}^{\text{emu}}) / \sigma_{V_X}^{\text{ref}}$ (dotted red), where $\sigma_{V_X}^{\text{ref}}$ is the standard deviation from the reference set and $\sigma_{V_X}^{\text{emu}}$ is the emulated dataset's one.

(dotted red lines), where $\sigma_{S_X}^{\text{emu}}$ is the standard deviation in the emulated dataset. We note that the average emulated S_4 coefficients closely follow the reference set, suggesting that this coefficient is among the most dominant ones in driving the gradient descent. Moreover, S_4 is the coefficient showing the most deviation from the Gaussian dataset, suggesting that S_4 is also the main driver

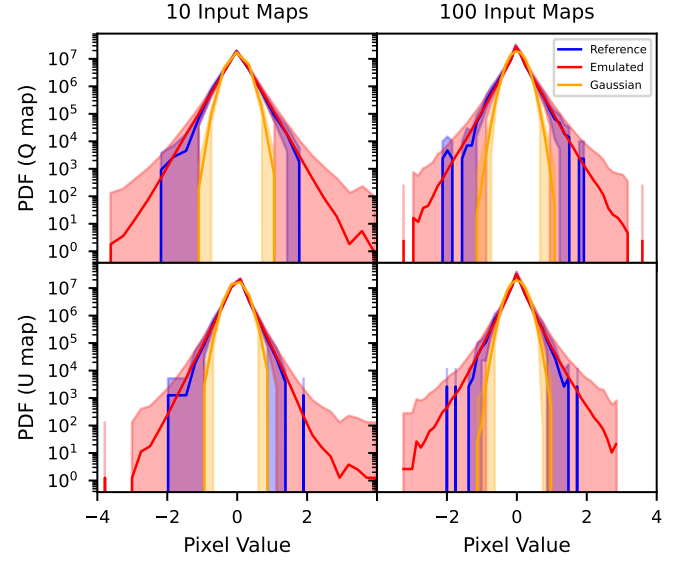


Fig. C.3. PDFs in pixel space for the maps in the reference (blue), emulated (red) and Gaussian (orange) datasets, for both the 10-maps (left panels) and 100-maps (right panels) input datasets. Top panels show the Q maps, while bottom ones the U maps. The shaded area represents the standard deviation of the maps PDFs.

behind the success of the generative model over the naive Gaussian augmentation from the angular power spectrum of the input maps. On the other hand, S_3 appears to contain limited information about the field under study: it is very small in amplitude and weakly constrained in this specific case, oscillating widely but still within the $1\text{-}\sigma$ band.

In the 10-input maps scenario (left panels), the SC coefficient variance at the smallest-convolved scales is slightly underestimated for all but S_3 , whereas in the 100-input maps case it is overestimated (again with the exception of S_3). We attribute this behavior to the different targets used during gradient descent, namely the average-target in the 10-maps setup versus the single-target in the 100-maps setup. Specifically, employing a single-target strategy overestimates the SC coefficient variance by incorporating both the “within-batch” variance introduced by mean-field gradient descent for each sampled target, and the “within-target” variance explained by the different targets in the input dataset. Despite this additional variance in the SC coefficients, the other key metrics presented in Section 5 are unaffected by the overestimation. Nonetheless, we expect that this excess variance could be mitigated by adding a suitable term in the loss function to penalize large “within-batch” variance. We leave further exploration of this approach to future work.

We compute the four Minkowski functionals V_1 , V_2 , V_3 and V_4 of the polarization maps (Fig. C.2) in the SO(3) formalism, which is the most appropriate for Q and U polarization maps (Carrón Duque et al. 2024). For this purpose, We use the `pynkowski` software¹⁷ (Carones et al. 2024; Carrón Duque et al. 2024). Minkowski functionals (hereafter MFs) provide a compact and computationally inexpensive summary of geometrical and topological information of the field (such as area, boundary length and Euler characteristic or genus as a function of a threshold) and are sensitive to long-range properties of the field and higher-order correlations not directly constrained by the power spectrum and SC coefficients. MFs represent

¹⁷ <https://github.com/javicarron/pynkowski>

therefore a complementary summary statistics, and demonstrating that the emulator reproduces the MFs of the reference set is thus a stringent test for our intended applications. Similarly to the SC coefficients, we quantify differences between the Minkowski functionals of the reference and emulated datasets, by defining their residuals as

$$\Delta V_X = \frac{\langle V_X^{\text{emu}} \rangle - \langle V_X^{\text{ref}} \rangle}{\sigma_{V_X}^{\text{ref}}},$$

where $X = 1, 2, 3, 4$ identifies the specific Minkowski functional, $\langle V_X^{\text{emu}} \rangle$ and $\langle V_X^{\text{ref}} \rangle$ are the average functionals in the emulated and reference sets, respectively, and $\sigma_{V_X}^{\text{ref}}$ is the standard deviation measured in the reference set. We also compare the variance in the two datasets by plotting

$$\Delta V_X^{\pm} = \frac{\langle V_X^{\text{emu}} \rangle - \langle V_X^{\text{ref}} \rangle \pm \sigma_{V_X}^{\text{emu}}}{\sigma_{V_X}^{\text{ref}}},$$

(dotted red lines), where $\sigma_{V_X}^{\text{emu}}$ is the standard deviation in the emulated dataset. The bottom panels of Fig. C.2 (solid red lines), show that ΔV_X remains well within the 1σ band across the threshold values u , although some residuals features are indeed prominent. Specifically, the feature around $u = 0$ (especially pronounced in V_0 in the $N_{\text{input}} = 10$ case and V_2 for $N_{\text{input}} = 100$) is a known effect of thresholding the MFs at such low-values, making the variance extremely small in the neighbourhood of $u = 0$ and therefore creating a spike in the ratio ΔV_X . Moreover, the MFs are higher-order, nonlinear statistics and therefore converge more slowly for any finite ensemble: this might be the cause of the pronounced wiggles in the residuals.

We also show the maps PDFs in pixel space for the reference (blue), emulated (red) and Gaussian (orange) datasets, for both the 10-maps (left panels) and 100-maps (right panels) input datasets (Fig. C.3). The top panels show the Q maps, while bottom ones the U maps. Again, the shaded area represents the standard deviation of the maps PDFs. Also here the emulated and reference maps appear to be remarkably consistent.

Appendix D: A note on the reference and input datasets

In this appendix, we offer additional clarification and motivation for the decision to construct the reference and the two input datasets by drawing Gaussian random realizations from the pixel-to-pixel covariance matrix of the original dataset produced from the TOD simulations (i.e. the reference covariance matrix described in Section 2), rather than using the original maps directly.

First, the empirical covariance matrix built from a finite dataset of simulations is plagued by sample variance. While this estimator is unbiased, the inverse of the empirical covariance – used for example in the χ^2 computation (such as the one in Eq. 8) and in likelihoods – is biased (Hartlap et al. 2007). Although in principle techniques such as the Hartlap factor (Hartlap et al. 2007) or the Ledoit–Wolf correction (Ledoit & Wolf 2003) can mitigate this, their performance can be difficult to evaluate, and they require additional modeling assumptions. Because one of our key metrics for assessing the performance of the emulator is its ability to reproduce the covariance matrix itself (alongside the χ^2 in pixel space), it is crucial for us to have access to a ground-truth reference covariance matrix. Therefore, instead

of emulating directly the original maps, we generate a new reference dataset—together with two input datasets of 10 and 100 maps, respectively—using the covariance matrix of the original dataset. This ensures that the reference covariance matrix serves as the ground truth for all these datasets.

Second, the probability density function of each pixel in the original maps is already close to Gaussian, due to the way these maps were generated (i.e. adding random Gaussian fluctuations at the TOD level; see Section 2). In our specific case, the pixel-to-pixel spatial information in the original dataset appears to be very well encoded in the reference covariance matrix: Gaussian realizations drawn from this covariance yield both maps and summary statistics (such as power spectrum, PDFs, Minkowski functionals and SC coefficients) very compatible with those of the original dataset. Therefore, no substantial difference in our results is expected when using the original dataset in lieu of the Gaussian realizations.

Third, generating Gaussian realizations from the reference covariance matrix enables us to create datasets that are completely independent, in a statistical sense, from the simulations used to estimate that same reference covariance. In high-accuracy analyses, even minimal overlap between the simulations used to build the covariance matrix and those considered as “data” (see for example Eq. 8) can, in principle, artificially suppress sampling variance and, consequently, the bias in the inverse covariance estimator. Our approach avoids this potential problem without having to split the original dataset into a set used for covariance matrix estimation and a set used only for validation, allowing us to use all 10,000 simulations in the original dataset to estimate the reference covariance.