

# FAST EMULATION OF ANISOTROPIES INDUCED IN THE COSMIC MICROWAVE BACKGROUND BY COSMIC STRINGS

M. A. PRICE<sup>†1</sup>, M. MARS<sup>1</sup>, M. M. DOCHERTY<sup>1</sup>, A. SPURIO MANCINI<sup>1</sup>, A. MARIGNIER<sup>1,2</sup>, AND J. D. McEWEN<sup>1,3</sup>

<sup>1</sup>Mullard Space Science Laboratory (MSSL), University College London (UCL), Holmbury St Mary, Dorking, Surrey RH5 6NT, UK

<sup>2</sup>Department of Earth Sciences, UCL, London, UK and

<sup>3</sup> Alan Turing Institute, Euston Road, London NW1 2DB, UK

*submitted XXX; accepted YYY*

## ABSTRACT

Cosmic strings are linear topological defects that may have been produced during symmetry-breaking phase transitions in the very early Universe. In an expanding Universe the existence of causally separate regions prevents such symmetries from being broken uniformly, with a network of cosmic string inevitably forming as a result. To faithfully generate observables of such processes requires computationally expensive numerical simulations, which prohibits many types of analyses. We propose a technique to instead rapidly *emulate* observables, thus circumventing simulation. Emulation is a form of generative modelling, often built upon a machine learning backbone. End-to-end emulation often fails due to high dimensionality and insufficient training data. Consequently, it is common to instead emulate a latent representation from which observables may readily be synthesised. Wavelet phase harmonics are an excellent latent representations for cosmological fields, both as a summary statistic and for emulation, since they do not require training and are highly sensitive to non-Gaussian information. Leveraging wavelet phase harmonics as a latent representation, we develop techniques to emulate string induced CMB anisotropies over a 7.2° field of view, with sub-arcminute resolution, in under a minute on a single GPU. Beyond generating high fidelity emulations, we provide a technique to ensure these observables are distributed correctly, providing a more representative ensemble of samples. The statistics of our emulations are commensurate with those calculated on comprehensive Nambu-Goto simulations. Our findings indicate these fast emulation approaches may be suitable for wide use in, *e.g.*, simulation based inference pipelines. We make our code available to the community so that researchers may rapidly emulate cosmic string induced CMB anisotropies for their own analysis. 

**Keywords:** Data Methods – methods: statistical – cosmology: miscellaneous – software: simulations

## 1. INTRODUCTION

Cosmic strings are linear topological defects produced when the Universe undergoes certain symmetry-breaking phase transitions, arising for example in a range of attempts at Grand Unification; for reviews see [Brandenberger 1994](#); [Vilenkin & Shellard 1994](#); [Hindmarsh & Kibble 1995](#); [Copeland & Kibble 2009](#). In an expanding Universe, the existence of causally separate regions prevents the symmetry from being broken in the same way throughout space, with a network of cosmic strings inevitably forming as a result ([Kibble 1976](#)). Cosmic strings are thus a well-motivated extension of the standard cosmological model and, while a string network cannot be solely responsible for the observed anisotropies of the cosmic microwave background (CMB) (since they could not explain the acoustic peaks of the CMB power spectrum; [Pen et al. 1997](#)), they could induce an important subdominant contribution.

The amplitude of any CMB anisotropies induced by cosmic strings is related to the string tension  $G\mu$ , where  $G$  is Newton's gravitational constant and  $\mu$  is the energy

per unit length of the string. In turn, the energy scale  $\eta$  of the string-inducing phase transition is directly related to  $\mu$  by  $\mu \sim \eta^2$ . Detecting signatures of cosmic strings would therefore provide a direct probe of physics of phase transitions in the early Universe at extremely high energy scales. Consequently, there has been a great deal of interest in constraining cosmic strings using observations of the CMB. In the majority of such analyses, signatures of string observables must be simulated, which is highly challenging.

Simulating accurate observable effects of a network of cosmic strings is a rich and highly computationally demanding field of research ([Albrecht & Turok 1989](#); [Bennett & Bouchet 1989, 1990](#); [Allen & Shellard 1990](#); [Hindmarsh 1994](#); [Bouchet et al. 1988](#); [Vincent et al. 1998](#); [Moore et al. 2002](#); [Landriau & Shellard 2003](#); [Ringeval et al. 2007](#); [Fraisse et al. 2008](#); [Landriau & Shellard 2011](#); [Blanco-Pillado et al. 2011](#); [Ringeval & Bouchet 2012](#)). There is an ongoing disagreement between Nambu-Goto (*e.g.* [Ringeval & Bouchet 2012](#)) and Abelian Higgs (*e.g.* [Hindmarsh et al. 2017](#)) simulation models regarding the decay of loops in string networks. In any case, in both models large-scale numerical simulations are required to

<sup>†</sup> E-mail: m.price.17@ucl.ac.uk

faithfully evolve string networks and simulate their observational effects. For example, the simulation of a single full-sky Nambu-Goto string-induced CMB map at sub-arcminute angular resolution can require in excess of 800,000 CPU hours, which is only possible by massively parallel ray tracing through thousands of Nambu-Goto string simulations (Ringeval & Bouchet 2012).

A variety of methods have been developed to search for string-induced contributions to the CMB, including power-spectrum constraints (Lizarraga et al. 2014a,b, 2016; Charnock et al. 2016), higher-order statistics such as the bispectrum (Planck Collaboration XXV 2014; Regan & Hindmarsh 2015) and trispectrum (Fergusson et al. 2010), and approaches such as edge detection (Lo & Wright 2005; Amsel et al. 2008; Stewart & Brandenberger 2009; Danos & Brandenberger 2010), Minkowski functionals (Gott et al. 1990; Ducout et al. 2013), wavelets and curvelets (Starck et al. 2004; Hammond et al. 2009; Wiaux et al. 2010; Planck Collaboration XXV 2014; Hergt et al. 2017; McEwen et al. 2017), level crossings (Sadegh Movahed & Khosravi 2011), peak-peak correlations (Movahed et al. 2013) and Bayesian inference (McEwen et al. 2017; Ciua & Hernández 2017). More recently, machine learning techniques have also been developed and shown great effectiveness (Ciua et al. 2019; Ciua & Hernández 2019, 2020; Vafaei Sadr et al. 2018; Torki et al. 2022). Due to the discrepancies between string simulation models, current constraints on the string tension depend on the model and simulation technique adopted. We avoid surveying the various constraints that have been reported in the literature to date and merely remark that typical constraints bound the string tension by  $G\mu \lesssim 10^{-7}$  (*e.g.* Planck Collaboration XXV 2014).

A critical component of all approaches to search for a cosmic string contribution in the CMB is the accurate simulation of string-induced CMB anisotropies. The massive computational cost of accurate string simulations, irrespective of the string simulation model, limits the effectiveness of cosmic string searches. This massive computational cost is currently unavoidable if the string network is to be accurately evolved and observables simulated faithfully. Compounding this, since strings induce significant contributions to CMB anisotropies at small angular scales, observables must be simulated at high-resolution. These limitations motivate alternative machine learning-based *emulation* techniques to generate realisations of synthetic observables, without the prohibitive computational overhead of full physical simulations, which is the focus of this article. Emulation is closely related to generative modelling and borrows many of the core ideas; naturally, many emulation methods leverage modern machine learning models, *e.g.* variational auto-encoders (Kingma & Welling 2013).

While techniques to emulate cosmic string-induced CMB anisotropies accurately do not exist currently, as far as we are aware, approaches to emulate other cosmological fields, such as large-scale structure, have been considered. Generative adversarial networks (Rodriguez et al. 2018; Mustafa et al. 2019; Perraudin et al. 2021; Feder et al. 2020) and variational auto-encoders (Chardin et al. 2019) have found some success emulating density fields directly (Piras et al. 2023). However, such end-to-end approaches are limited to low to moderate dimensions

and require large volumes of training data. To circumvent the issues of high dimensionality and large volumes of training data, an alternative approach is to emulate some latent representation from which observables may be readily synthesised. For example, it is common to first emulate a power spectrum, *e.g.* through polynomial regression (Jimenez et al. 2004; Fendt & Wandelt 2007), Gaussian processes (Heitmann et al. 2009; Lawrence et al. 2010; Ramachandra et al. 2021; Euclid Collaboration et al. 2021), or multilayer perceptrons (Auld et al. 2008; Agarwal et al. 2012; Bevins et al. 2021; Spurio Mancini et al. 2022), from which Gaussian realisations may trivially be generated. For the emulation of string-induced anisotropies, which are highly non-Gaussian, adopting the power spectrum as a latent representation is not well-suited.

In this article we propose a technique to emulate CMB anisotropies induced by networks of cosmic strings that both eliminates the computational bottleneck and captures non-Gaussian structure. Our emulation technique adopts the recently developed wavelet phase harmonics (Mallat et al. 2020; Allys et al. 2020; Zhang & Mallat 2021; Brochard et al. 2022), a form of second generation scattering transform (Mallat 2012; Bruna & Mallat 2013), as a latent representation. Once a wavelet phase harmonic representation is computed from a *small ensemble* of physical simulations, our approach can then be used to rapidly generate high-resolution realisations of the cosmic string induced CMB anisotropies in under a minute, starkly contrasting the computational cost of a single simulation. Such an acceleration unlocks a variety of analysis techniques, including but not limited to those which necessitate the repeated synthesis of observables, *e.g.* Bayesian inference which often relies on sampling. In particular our approach is suitable for use in simulation based inference (SBI) pipelines (Cranmer et al. 2020; Spurio Mancini et al. 2022), where the likelihood is either not available or too costly to be evaluated, and inference relies solely on the ability to efficiently simulate or emulate observables. Such techniques are predicated on the ability to generate observations that are not only realistic but are also correctly distributed. In this article we explore this second qualification as well, which is often overlooked despite being critical for scientific studies.

The remainder of this article is structured as follows. In Section 2 we provide an overview of generative modelling within the context of cosmology. We then present our approach for the rapid emulation of cosmic string induced CMB anisotropies in Section 3, which we subsequently validate in Section 4. Finally, we discuss the impact of these results and draw conclusions in Section 5.

## 2. GENERATIVE MODELLING OF PHYSICAL FIELDS

Generative modelling is a term broadly ascribed to the generation of synthetic observables that approximate authentic observables. Throughout the following discussion we will refer to authentic observables by  $\mathbf{x}_{\text{True}}$  and synthetic observables by  $\mathbf{x}_{\text{Syn}}$ , which can be either simulated or emulated observables, denoted  $\mathbf{x}_{\text{Sim}}$  and  $\mathbf{x}_{\text{Emu}}$  respectively. A diverse range of generative models exist with varying motivations, although many are motivated by the *manifold hypothesis* (Bengio et al. 2013).

**Manifold Hypothesis:** *A given authentic observable  $\mathbf{x}_{\text{True}} \in \mathcal{X}$ , where  $\mathcal{X}$  is the ambient space with dimensionality  $d_{\mathcal{X}}$ , is hypothesised to live on a manifold  $\mathcal{S} \subseteq \mathcal{X}$  with dimensionality  $d_{\mathcal{S}} \leq d_{\mathcal{X}}$ , embedded within  $\mathcal{X}$ .*

Intuitively, this becomes apparent by considering natural images and making the following realisations. First, images generated by uniformly randomly sampling each pixel are extremely unlikely to be meaningful (Pope et al. 2021). Secondly, images are highly locally connected through various transformations (*e.g.* contrast, brightness), symmetries (*e.g.* translations, scaling), and diffeomorphisms (one-to-one invertible mappings, *e.g.* stretching). There is strong evidence to suggest the manifold hypothesis is correct (Bengio et al. 2013), with algorithmic verification by Fefferman et al. (2016). In any case, where additional flexibility is necessary a union of manifolds hypothesis may be adopted with similar justification (Brown et al. 2022).

For a complete description of the generative model one must also characterise the data generating distribution on this manifold, *i.e.* the likelihood with which any given synthetic observable is to have been observed. In such a case one may interpret  $\mathcal{S}$  as a *statistical manifold* (see *e.g.* Amari 2016; Nielsen 2020).

**Statistical Manifold:** *A manifold  $\mathcal{S}$  on which observables  $\mathbf{x}_{\text{True}} \in \mathcal{S}$  live that is endowed with a probability distribution  $\mathbb{P}_{\text{True}}$ .*

Under the statistical manifold hypothesis the generative problem is two-fold: (i) how best to generate realistic synthetic observables  $\mathbf{x}_{\text{Syn}} \in \mathcal{S}$ , and (ii) how to ensure the probability distribution  $\mathbb{P}_{\text{Syn}}$  of  $\mathbf{x}_{\text{Syn}}$  matches  $\mathbb{P}_{\text{True}}$ . That is, how best to not only approximate the embedded manifold but also the distribution over that manifold. With machine learning techniques problem (i) can often be satisfied, provided access to a sufficiently large bank of data  $\mathbf{d}$ . However problem (ii) is less straightforward to address and in many cases depends on the degree to which the distribution of  $\mathbf{d}$  traces  $\mathbb{P}_{\text{True}}$ . It should be noted that, attempting to model both  $\mathcal{S}$  and  $\mathbb{P}_{\text{True}}$  with maximum-likelihood based methods can be pathological when the ambient dimensionality of the space  $\mathcal{X}$  is significantly different to that of  $\mathcal{S}$  (Dai & Wipf 2019). At a high-level this effect, which is referred to as *manifold overfitting*, occurs when the manifold  $\mathcal{S}$  is learned but the distribution over  $\mathcal{S}$  is not (Loaiza-Ganem et al. 2022).

One way in which this pathology may be solved is by first learning the data-distribution on a latent representation (equivalently a summary statistic) with low dimensionality (ideally equal to that of  $\mathcal{S}$ ) before decoding to an approximation of the data-distribution. This approach to learning the data distribution was first explored by Loaiza-Ganem et al. (2022), who show that if the latent representation is a generalized autoencoder, then the data-distribution on  $\mathcal{S}$  may be recovered theoretically (see Loaiza-Ganem et al. 2022, Theorem 2). A variety of other effective methods have been proposed to handle this pathology (Arjovsky et al. 2017; Horvat & Pfister 2021; Song & Ermon 2019; Song et al. 2020).

The importance of the above criteria when generating natural images or physical fields differs greatly. In most applications, it is sufficient to rapidly generate inexpensive synthetic observables with high fidelity. For exam-

ple, in the large-scale generation of synthetic natural images or celebrity faces Rombach et al. (2022), matching the correct data generating distribution  $\mathbb{P}_{\text{True}}$  is perhaps less important. For scientific analysis, however it is typically necessary to generate synthetic observables that not only live on or in the neighbourhood of  $\mathcal{S}$ , but also are approximately drawn from  $\mathbb{P}_{\text{True}}$ . An accurate approximation of the distribution on the manifold is critical for use in, for example, simulation based inference pipelines.

## 2.1. Simulation

Many generative models have been developed for a broad range of applications, however in this article we will consider two categories: simulation and emulation. From the perspective of a cosmologist, simulation entails the time evolution of initial conditions, *e.g.* an initial field  $\mathbf{x}_0$ , governed by cosmological parameters  $\boldsymbol{\theta}$ , to some late-universe observables  $\mathbf{x}$ . Such evolution is designed to model the underlying physics of a universe from the grandest to smallest scales, which can become incredibly complex and non-linear (Hockney & Eastwood 2021). Extracting information at higher angular resolutions is of increasing importance as recent and forthcoming cosmological experiments probe smaller scales with greater sensitivity. Simulating small-scale physics is therefore critical, necessitating high resolution simulations to faithfully represent late-universe observables, which is highly computationally demanding. Computationl hurdles aside, it is important to note that, provided the core physics is sufficiently captured, an ensemble of simulated observables will reliably trace  $\mathbb{P}_{\text{True}}$ , which is critical for subsequent analyses.

**Simulation:** *A generative model which directly encodes the dynamics of a physical system, evolving some initial conditions over time to a late universe observable  $\mathbf{x}_{\text{Sim}}$ . The dynamics of a system are governed by parameters  $\boldsymbol{\theta}$ .*

Such generative models are dependent only on an understanding of both the initial conditions, parameters  $\boldsymbol{\theta}$ , and the underlying physics, and do not need to model the statistical distribution of the data directly since it is captured implicitly by the simulation process.

## 2.2. Emulation

One may instead emulate observations, circumventing simulation entirely by approximating a mapping from cosmological parameters  $\boldsymbol{\theta}$  to synthetic late-universe observables  $\mathbf{x}_{\text{Emu}}$ . Provided training data  $\mathbf{d} = \{\boldsymbol{\theta}, \mathbf{x}_{\text{True}}\}$  one may attempt to train a model to approximate this mapping directly. End-to-end approaches are reliant on a sufficiently large volume of training data, the amount of which scales with both dimensionality and functional complexity. Cosmology is fundamentally restricted to synthetic training data, which can only be accurately and reliably generated through computationally expensive simulations. While generating small numbers of such simulations is expensive but achievable (Nelson et al. 2019; Villaescusa-Navarro et al. 2020), generating large ensembles of such simulations is often simply not feasible.

Consequently, to ameliorate these concerns it is common to instead emulate a compressed latent representation from which observables may readily be synthesised. In the following we define a compression  $\Phi : \mathbf{x} \mapsto \mathbf{z} \in \mathcal{Z}$ ,

where  $\mathcal{Z}$  is of dimension  $d_{\mathcal{Z}}$ . Further, consider the setting where we constrain the ratio  $r = d_{\mathcal{Z}}/d_{\mathcal{X}} < 1$ , such that  $\mathbf{z}$  is a potentially lossy compressed representation of  $\mathbf{x}$ . The objective is therefore to first approximate the latent mapping  $\Lambda : \boldsymbol{\theta} \mapsto \mathbf{z}_{\text{Emu}}$  from which observables may be synthesised by taking into consideration the latent compression  $\mathbf{z}_{\text{Emu}} = \Phi(\mathbf{x}_{\text{Emu}})$ . To learn an approximation of  $\Lambda$  requires less training data due to the reduction in dimensionality. Hence, a trade-off between the complexity of  $\Lambda$  and  $\Phi$  exists and so one can balance between data requirements and the information lost during compression. As the compression ratio  $r$  decreases, *i.e.* greater compression, the data requirements diminish, however conversely the compression loss is likely to increase.

Popular summary statistics such as the power-spectrum are emulated in this manner, from which Gaussian realisations may be generated trivially (see *e.g.* Auld et al. 2008; Agarwal et al. 2012; Bevins et al. 2021; Spurio Mancini et al. 2022). However, the power spectrum is a particularly ill-suited latent representation for the synthesis of cosmic string induced CMB anisotropies, which are highly non-Gaussian in nature. Hypothetically, one could adopt a variational auto-encoder (Kingma & Welling 2013) as an effective latent representation; in fact Loaiza-Ganem et al. (2022) have recently had some success in this regard. It is reasonable to presume such an approach would be sensitive to non-Gaussian information, however for aforementioned reasons gathering sufficient training data is infeasible. This dichotomy therefore motivates the development of latent representations that are sensitive to non-Gaussian information and do not require substantial training data.

**Latent emulation:** *A two-step generative model, including a mapping  $\Lambda$  from cosmological parameters  $\boldsymbol{\theta}$  to latent variables  $\mathbf{z}_{\text{Emu}}$ , from which observables  $\mathbf{x}_{\text{Emu}}$  are synthesised given knowledge of the compression mapping  $\Phi$  that maps from observables to the latent space, i.e.  $\mathbf{z}_{\text{Emu}} = \Phi(\mathbf{x}_{\text{Emu}})$ .*

The reduced dimensionality of  $\mathbf{z}_{\text{Emu}}$  alleviates training data requirements, introducing a trade-off between the complexity of the mapping and the compression loss, which can effect the quality of synthesis. Since  $\Phi$  need only be surjective (onto), there typically exists some variability in synthetic observables, as potentially many observables correspond to a single latent vector. However, this implicit variability is by no means guaranteed to match the data generating distribution  $\mathbb{P}_{\text{True}}$  on  $\mathcal{S}$ . One should note that in the setting of Loaiza-Ganem et al. (2022), where  $\Phi$  is a generalized autoencoder, provided  $d_{\mathcal{Z}} = d_{\mathcal{S}}$ , and the distribution on  $\mathcal{Z}$  is sufficiently captured, the induced distribution  $\mathbb{P}_{\text{Syn}}$  recovers  $\mathbb{P}_{\text{True}}$  to a good approximation. Such an approach is appropriate for computer vision tasks, where data is far from a limiting factor. However, for cosmological applications insufficient data is available to learn such latent representations, motivating the adoption of designed representations, *e.g.* wavelet-based representations.

### 2.3. Wavelet Phase Harmonics

The wavelet phase harmonics (WPH) are a form of second generation scattering transform (Mallat 2012; Ally et al. 2019; Mallat et al. 2020) which can be directly contrasted with convolutional neural networks. For WPHs

filters are defined by wavelets rather than learned in a data-driven manner. Drawing inspiration again from machine learning, once a signal of interest has been convolved with the wavelet of a given scale, point-wise nonlinearities are applied through the phase harmonic operator  $\mathbf{w} \mapsto [\mathbf{w}]^p = |\mathbf{w}| \cdot e^{ip\arg(\mathbf{w})}$ , which is simply a rotation of some complex vector  $\mathbf{w}$ . As such, rotations induce magnitude and scale independent non-linearities, hence spatial information may be synchronised across scales, from which moments (covariances between distinct convolutions) are computed. Consequently WPH provide a latent representation particularly well suited for spatially homogeneous images, *e.g.* textures (Zhang & Mallat 2021). Furthermore, WPH can be shown to be highly sensitive to non-Gaussian information (Portilla & Simoncelli 2000), making them ideal latent representations for cosmic string induced CMB anisotropies.

WPH and their predecessors, the first generation wavelet scattering transform, have successfully been applied to probe weak gravitational lensing (Cheng et al. 2020; Cheng & Ménard 2021; Valogiannis & Dvorkin 2022; Eickenberg et al. 2022), the removal of non-Gaussian foreground contaminants (Allys et al. 2019; Regaldo-Saint Blanchard et al. 2020, 2021; Jeffrey et al. 2022), classification of magnetohydrodynamical simulations (Saydjari et al. 2021), and exploration of the epoch of reionisation (Greig et al. 2022; Lin et al. 2022). Many of these applications have adopted the WPH as a latent representation from which realistic observables have been emulated. However, as far as we are aware, to date little consideration has been given to the probability distribution of such observables (see Section 2).

There are two distinct way to construct maximum entropy generative models, these being the micro- and macro-canonical approaches, which relate to the associate ensembles in statistical physics. We have discussed the micro-canonical case, wherein new realizations which have the same latent representation are iteratively generated, and provided an argument to why such an approach can result in limited variability. In contrast to this, the macro-canonical case consists in explicitly constructing a probability distribution for which the WPH are not fixed. This probability distribution can in turn be related to the physical Hamiltonian of the process under study, however the difficulty is then how one samples from this ensemble (Marchand et al. 2022).

## 3. FAST EMULATION OF COSMIC STRING INDUCED ANISOTROPIES

Having outlined both generative modelling and latent emulation in the context of physical fields, we next describe how these concepts can be leveraged to rapidly generate realisations of late-universe cosmic string induced anisotropies in the CMB.

First let us explicitly formulate our emulation problem following the notation of Section 2.2. We seek to emulate late-universe string induced anisotropies  $\mathbf{x}_{\text{Emu}}$  from cosmological parameters  $\boldsymbol{\theta}$ . Fortunately, in the case of cosmic string induced CMB anisotropies there is only a single parameter  $\boldsymbol{\theta} = G\mu$ , the string tension (Kibble 1976). Moreover, the observed anisotropies transform trivially under  $\mu \rightarrow \mu'$ , specifically this transformation is simply a scaling  $\mathbf{x}_{\text{Emu}} \rightarrow (\mu'/\mu)\mathbf{x}_{\text{Emu}}$ . Therefore, provided one is able to generate emulated observables for

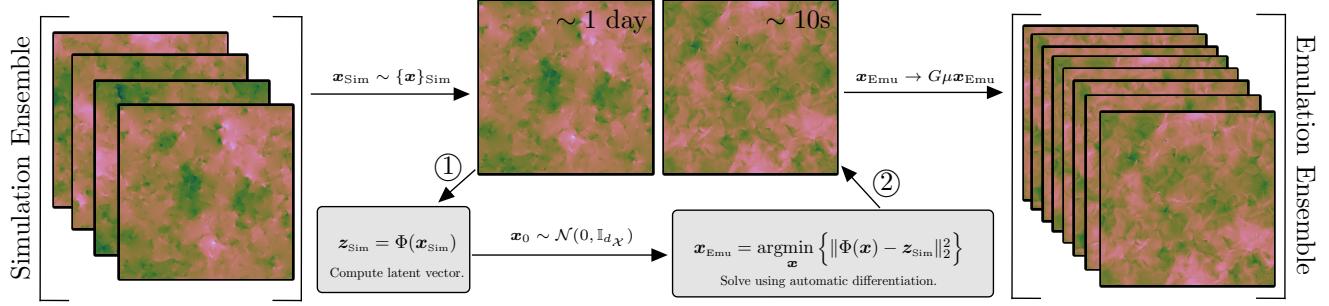


FIG. 1.— An overview of the process by which a small ensemble of simulated observations can be extremely augmented with emulated observations for arbitrary string tension  $G\mu$ . In step 1 (compression) we simply draw a uniform random simulation  $\mathbf{x}_{\text{Sim}}$  from which a reference latent vector  $\mathbf{z}_{\text{Sim}}$  is calculated. In step 2 (synthesis) we take a random Gaussian realisation  $\mathbf{x}_0$  and, using automatic differentiation of the compression mapping  $\Phi$  to iteratively minimise a standard  $\ell_2$ -loss function, recovering solutions  $\mathbf{x}_{\text{Emu}}$  such that  $\Phi(\mathbf{x}_{\text{Emu}}) = \mathbf{z}_{\text{Sim}}$ . Steps 1 and 2 can straightforwardly be repeated many times, generating an ensemble of emulated maps which can be (potentially much) larger than the small collection of simulated observables. In this way this approach may be thought of as extreme data augmentation.

a single string tension, it is straightforward to generate them for all string tensions. As such, in the following we simplify to a single fixed  $\mu$  from which all  $\mu' \neq \mu$  can readily be generated *a posteriori*.

We consider how to robustly synthesise string induced anisotropies  $\mathbf{x}_{\text{Emu}}$  from their WPH representation  $\mathbf{z}_{\text{Emu}}$ . More formally, for a given reference latent vector  $\mathbf{z}_{\text{Emu}}$ , which we condition on, we efficiently synthesise observations  $\mathbf{x}_{\text{Emu}}$  which satisfy  $\Phi(\mathbf{x}_{\text{Emu}}) = \mathbf{z}_{\text{Emu}}$ . Additionally, we provide a strategy by which ensembles of such emulated observables can, at least approximately, be shown to be distributed appropriately, *i.e.*  $\mathbb{P}_{\text{Emu}} \approx \mathbb{P}_{\text{True}}$ . To this end we leverage a small set of simulated observables  $\mathbf{x}_{\text{Sim}}$  as a trellis, upon which our emulation process grows.

Throughout this work we will adopt WPHs as our compressed latent representation  $\Phi$  (Mallat et al. 2020), which is highly sensitive to non-Gaussian information (Portilla & Simoncelli 2000), is numerically efficient to evaluate, and does not require training data since it adopts designed rather than learned filters. We make use of the GPU-accelerated PyTorch package PyWPH<sup>1</sup> which implements the transform discussed in Regaldo-Saint Blanchard et al. (2021), and by default adopts bump steerable wavelets (Mallat et al. 2020).

### 3.1. Generating String Induced Anisotropies

In the following we work under the assumption that a (potentially very) limit number of simulated observables are available, from which we will generate arbitrarily many synthetic observables. Offline, we apply  $\Phi$  to compress this *training* set into latent vectors that we condition on during synthesis. We then iteratively emulate many observations  $\mathbf{x}_{\text{Emu}}$  such that  $\Phi(\mathbf{x}_{\text{Emu}}) \approx \mathbf{z}_{\text{Sim}}$  through gradient-based algorithms given an appropriate loss surface. Here we chose to minimise the standard Euclidean  $\ell_2$ -loss  $\mathcal{L}(\mathbf{x}) = \|\Phi(\mathbf{x}) - \mathbf{z}_{\text{Sim}}\|_2^2$ . To achieve this in practice requires software to calculate both the compression  $\Phi$  and necessary gradients, both of which are straightforwardly provided by PyWPH. An iterative approach, such as the one presented here, has also been adopted to successfully emulate a variety of cosmological signals, from density fields (Allys et al. 2019, 2020) to foreground contaminants (Regaldo-Saint Blanchard et al.

2021; Jeffrey et al. 2022).

For our current work we match the latent representation by maximum-likelihood estimation. One may instead perform maximum-a-posteriori estimation by enforcing regularity constraints. For example, cosmic string networks are close to piece-wise constant, hence emulation of their induced anisotropies may benefit from a total-variation norm regularisation (gradient sparsity), however we leave this exploration to a later date.

In this work we use the L-BFGS algorithm to minimise the loss function, which is a variant of the quasi-Newton method BFGS (Byrd et al. 1995) and typically require at most 100 iterations to converge to a solution  $\mathbf{x}_{\text{Emu}}$  for which the loss functions is below an acceptable tolerance. Visually, we confirm that these solutions  $\mathbf{x}_{\text{Emu}}$  display similar characteristics to those generated through comprehensive simulations, indicating that they live on, or in the neighbourhood of, the embedded manifold  $\mathcal{S}$ . In many cases generating visually realistic synthetic observables alone is sufficient, *e.g.* for natural images. However, to leverage these techniques for scientific inference it is important to ensure an ensemble of synthetic observables are distributed according to the data generating distribution  $\mathbb{P}_{\text{True}}$ .

### 3.2. Matching the Probability Distribution

Suppose a single simulation is available, from which  $m$  synthetic observables  $\{\mathbf{x}_{\text{Emu}}\}_m$  may readily be emulated. From the surjectivity of  $\Phi$  our emulated set of observables will exhibit some degree of variability, however this distribution is by no means guaranteed to match the true underlying data-generating distribution. In fact this is highly unlikely.

Were one to evaluate the expectation  $\mathbb{E}[\cdot]$  of a summary statistic of interest  $\Omega$  over these  $m$  emulated observables they are likely to approximate the point statistics of the single simulation, but may not match the summary statistics averaged over an ensemble of  $n$  simulations  $\{\mathbf{x}_{\text{Sim}}\}_n$ . This is to say that although our ensemble of emulated realisations sufficiently match a single simulation, they do not correctly characterise an ensemble of simulations  $\mathbb{E}[\{\Omega(\mathbf{x}_{\text{Emu}})\}_m] \neq \mathbb{E}[\{\Omega(\mathbf{x}_{\text{Sim}})\}_n]$ . Therefore such emulations are likely to bias any subsequent statistical analysis. An analogous argument may be made toward  $\text{Var}[\{\Omega(\cdot)\}_m]$  and other higher order descriptors.

<sup>1</sup> <https://github.com/bregaldo/pywph>

**Algorithm 1** Emulation of cosmic string signatures

---

First take our small set of  $m$  simulations  $\{\mathbf{x}_{\text{Sim}}\}_m$  and compute their latent vectors, which we will condition on during synthesis.

```
procedure GENERATE LATENT ENSEMBLE( $\Phi, \{\mathbf{x}_{\text{Sim}}\}_m$ )
  for  $i \in [0, m)$  do
     $\mathbf{z}_{\text{Sim}} = \Phi(\mathbf{x}_{\text{Sim}})$ 
  return  $\{\mathbf{z}_{\text{Sim}}\}_m$ 
```

Draw a uniform random reference latent vector  $\mathbf{z}_{\text{Sim}}$  upon which we will condition. Starting from white noise, use automatic differentiation to find  $\mathbf{x}_{\text{Emu}}$  such that  $\Phi(\mathbf{x}_{\text{Emu}}) \approx \mathbf{z}_{\text{Sim}}$ .

```
procedure EMULATE FIELD( $\mu, \Phi, \{\mathbf{z}_{\text{Sim}}\}_m$ )
   $\mathbf{x} \sim \mathcal{N}(0, \mathbb{I}_{d_x})$                                  $\triangleright$  Random initial field
   $\mathbf{z}_{\text{Sim}} = \{\mathbf{z}_{\text{Sim}}\}_{j \sim \mathcal{U}\{0, m-1\}}$            $\triangleright$  Draw latent vector
function LOSS FUNCTION( $\Phi, \mathbf{x}, \mathbf{z}_{\text{Sim}}$ )
   $\mathcal{L} = \|\Phi(\mathbf{x}) - \mathbf{z}_{\text{Sim}}\|_2^2$ 
  return  $\mathcal{L}, \frac{\partial \mathcal{L}}{\partial \mathbf{x}}$                                  $\triangleright$  Return loss and gradient.
 $\mathbf{x}_{\text{Emu}} = \text{L-BFGS-B}(\text{Loss Function}, \mathbf{x}, \mathbf{z}_{\text{Sim}})$ 
return  $G\mu \mathbf{x}_{\text{Emu}}$                                  $\triangleright$  Rescale to string tension  $\mu$ 
```

---

The solution we propose is to instead work with a small *training* ensemble of simulated observables, which more adequately represent the data-generating distribution. During subsequent statistical analysis whenever observables are required, a random latent representation is uniformly drawn from this training set and used to generate  $\mathbf{x}_{\text{Emu}}$  through the method outlined in Section 3.1.

In this way one may reasonably expect to find that the statistics computed from a set of emulated observables should match those computed on a set of simulated observables. That is to say that  $\mathbb{E}[\{\Omega(\mathbf{x}_{\text{Emu}})\}_m] \approx \mathbb{E}[\{\Omega(\mathbf{x}_{\text{Sim}})\}_n]$ , provided  $n$  and  $m$  are each sufficiently large. Increasing the amount of training data will improve the reliability and accuracy with which the distribution of our limited ensemble of training simulations matches the underlying data-generating distribution, improving the degree to which emulated observables are approximately drawn from the true data generating distribution  $\mathbf{x}_{\text{Emu}} \sim \mathbb{P}_{\text{True}}$ .

To summarise this approach makes the following assertion: the distribution over observables upon which we condition during emulation is transferred to the distribution of our emulated observables. Some augmentation is applied to this distribution, as there is some variability during synthesis, but typically this is a comparatively small effect. Hence, using a small training set of simulated observables provides a straightforward means by which the distribution of emulated observables can be made substantially more realistic.

**Emulation as Augmentation:** Our approach may be considered extreme data augmentation, wherein latent emulation bridges the gap between the number of simulations necessary for inference and those which may feasibly be generated. The limited span of our small ensemble of simulations is enhanced by the variability (expressiv-

ity) induced by the surjectivity of  $\Phi$ .

Alternatively, one may attempt to enhance the variability of synthetic observables by modelling a probability measure on the latent representation directly, as was promoted by Loaiza-Ganem et al. (2022). In the case where  $\Phi$  is a generalized autoencoder the compressive mapping is injective and learned. However, when  $\Phi$  is given by the WPHs it is not at all obvious which distribution over latent variables corresponds with  $\mathbb{P}_{\text{True}}$ . There are several approaches one may wish to consider however we leave this for future work (see e.g. De Bortoli et al. 2022).

### 3.3. Algorithm and Computational Efficiency

Our approach involves three primary steps: (1) A small training set of latent vectors is computed from simulations once; (2) A random latent vector  $\mathbf{z}_{\text{Sim}}$  is drawn from this ensemble; and (3) the loss discussed in Section 3.1 is minimised to generate an emulated observable such that  $\Phi(\mathbf{x}_{\text{Emu}}) \approx \mathbf{z}_{\text{Sim}}$ . These steps are outlined in Algorithm 1 and Figure 1, and are implemented in code which we make publicly available. <sup>2</sup>

We benchmarked the computational overhead for our approach on a single dedicated NVIDIA A100 Tensor Core GPU with 40GB of device memory. Compiling the PyWPH kernel, our compression  $\Phi$ , for  $1024 \times 1024$  images takes  $\sim 11$ s on average and occupies  $\sim 27$ GB of the available onboard memory; indicating the PyWPH software is fast but not yet memory efficient. It should be noted that we adopt default configuration of all PyWPHv1.0 hyper-parameters, and that subsequent PyWPH releases demonstrate further acceleration. Synthesis of a single string induced anisotropies takes 100 L-BFGS iterations with a wall-time of  $\mathcal{O}(100)$ s. In practice, the quality of synthetic observations degrades only slightly if the optimiser is run for significantly fewer iterations, and so the wall-time can easily be reduced to less than a minute. As a baseline; a single flat-sky Nambu-Goto simulation at this resolution takes more than a day of wall-time, and a full-sky simulation can take in excess of 800,000 CPU hours.

## 4. VALIDATION EXPERIMENTS

To demonstrate the efficacy of the emulation process discussed at length in Section 3 and summarised in Algorithm 1, we generate a set of synthetic cosmic string induced CMB anisotropies, the summary statistics of which are validated against those computed over an ensemble of state-of-the-art Nambu-Goto string simulations.

### 4.1. Nambu-Goto String Simulations

Due to the multiscale nature of wavelets, string induced CMB anisotropies may be emulated for a wide variety of string models, given a field simulation. In this analysis we adopt the Nambu-Goto string simulations of Fraisse et al. (2008), although in principle alternative string simulations could be considered.

These Nambu-Goto string induced anisotropies are convolved with a 1 arcminute observational beam in line

<sup>2</sup> <https://github.com/astro-informatics/stringgen>

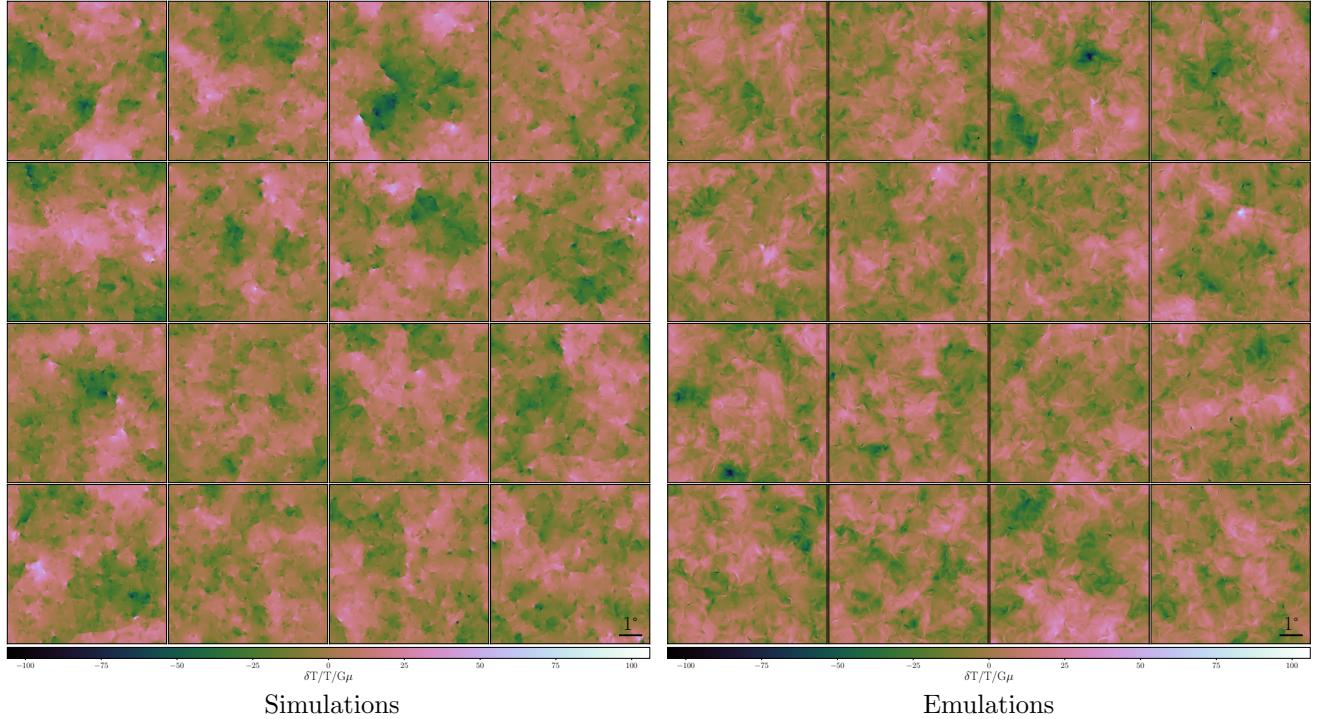


FIG. 2.— **Left:** A gallery of simulated Nambu-Goto cosmic string induced CMB anisotropies randomly sampled from an ensemble of 1000 of such images. Each of these simulations can take in excess of a day to compute. **Right:** A gallery of emulated string induced anisotropies, each of which take on average under a minute to generate, and are statistically indistinguishable from their simulated counterparts displayed on the left. These synthetic string induced anisotropies are emulated using the methods presented in this article.

with current ground based observations, such as the Atacama Cosmology Telescope (Louis et al. 2014) and South Pole Telescope (Chown et al. 2018). It is important to note that these simulated flat-sky maps are generated using discrete Fourier transforms and that the genuine cosmic string power spectrum goes as  $\sim 1/k$ . Therefore, these simulations introduce substantial aliasing at small scales. Such beam convolutions mitigate aliasing by removing any excess power at high frequencies. In total, we have 1000 state-of-the-art Nambu-Goto string maps, each of dimension  $1024 \times 1024$ , covering a  $7.202^\circ$  field of view at sub-arcminute resolution.

#### 4.2. Methodology

We partition the 1000 available  $1024 \times 1024$  Nambu-Goto string simulations into *training* and *validation* datasets, with 300 and 700 simulations respectively. For each simulation we compute the associated WPH representation, which we store for subsequent use. Note that we adopt the machine learning nomenclature for consistency, though training is not necessary since we adopt WPHs as our compression  $\Phi$ , which provide a designed rather than learned latent representation space. Following the method outlined in Algorithm 1, we generate 700 emulated string induced anisotropies, each time uniformly randomly sampling a set of WPH coefficients  $\mathbf{z}_{\text{Sim}}$  from the training set. Finally, we compute summary statistics over our emulated CMB anisotropies, which we validate against those computed over the validation dataset.

#### 4.3. Validation

A gallery of randomly selected simulated and emulated string induced anisotropies can be seen in Figure 2; the statistical properties of these maps appear very similar to the eye. Though it is necessary that emulated observables  $\mathbf{x}_{\text{Emu}}$  are of high fidelity, one must further ensure that an ensemble of such observables correctly characterise authentic CMB anisotropies. That is, emulated observables for scientific applications must be both of high fidelity and appropriate variability. This duality is discussed in Section 2. One must ensure that  $\mathbf{x}_{\text{Emu}}$  are, at least approximately, distributed according to the data generating distribution  $\mathbb{P}_{\text{True}}$ . If this second condition is not satisfied, although one may recover individual maps which appear reasonable, the aggregate statistics of such maps will likely be incorrect.

Two naïve approaches can help elucidate this point. Suppose one selects a single latent vector  $\mathbf{z}_{\text{Sim}}$  from which many synthetic observables are generated. We explored this and indeed find that the statistics of these anisotropies highly concentrate around the point statistics associated with our chosen latent vector  $\mathbf{z}_{\text{Sim}}$  and do not fully capture  $\mathbb{P}_{\text{True}}$ . Suppose instead one attempts to ameliorate this by constructing an averaged latent representation  $\mathbb{E}[\{\mathbf{z}_{\text{Sim}}\}_k]$  over  $k$  training simulations, from which many synthetic observables are generated. Again, we explored this and find that the statistics highly concentrated around the mean latent vector and do not remotely capture  $\mathbb{P}_{\text{True}}$ . However, it should be noted that cosmic string induced anisotropies exhibit structure which is particularly difficult to model, so it may be that such approaches are sufficient for other applications.

To ensure we capture  $\mathbb{P}_{\text{True}}$  sufficiently to support the

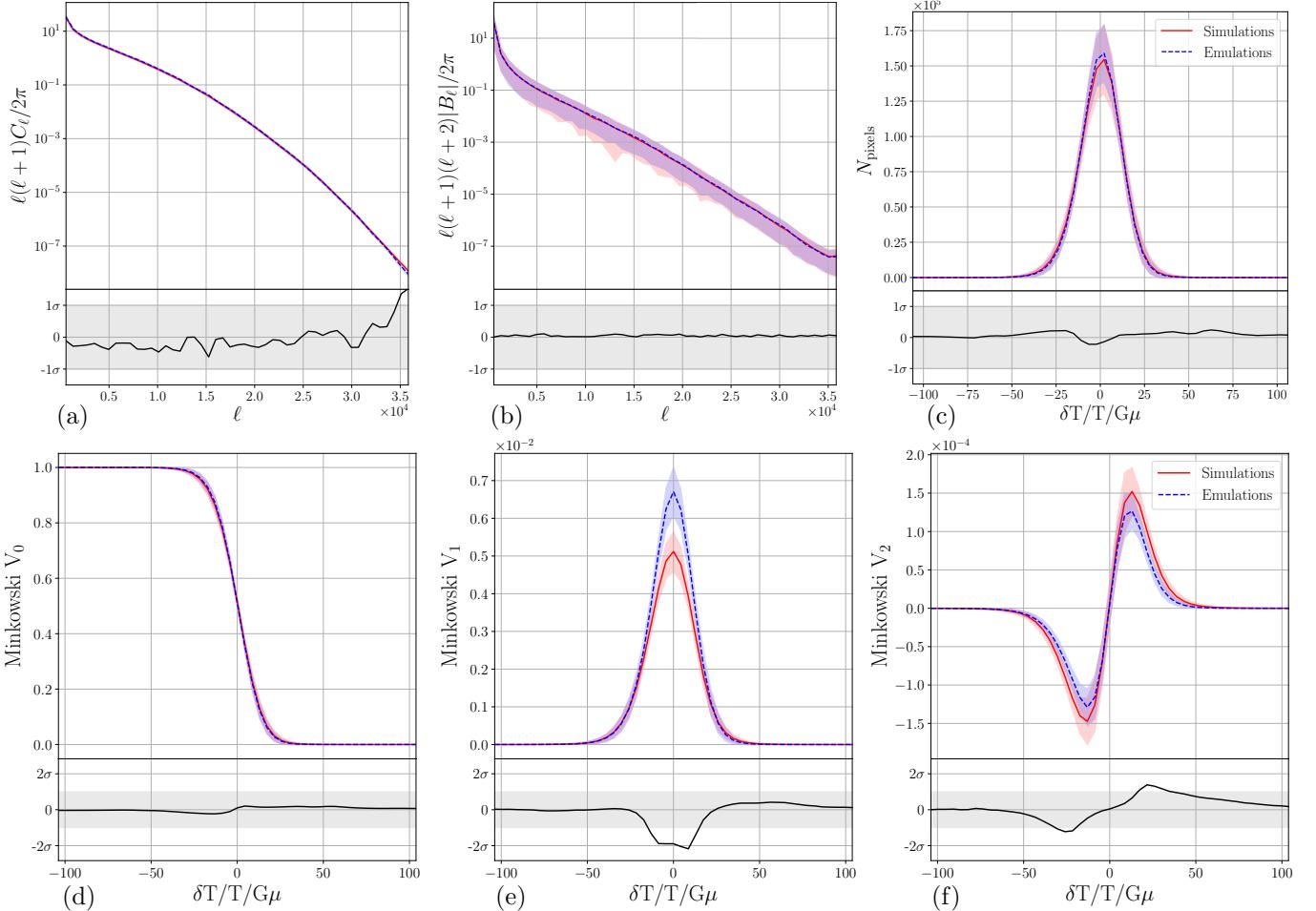


FIG. 3.— Summary statistics considered for the validation of the emulation techniques presented in this article. Each panel displays the mean of a summary statistic (line) and its variance ( $1\sigma$ , shaded), for 700 simulated (red, solid) and emulated (blue, dashed) string induced CMB anisotropies. At the bottom of each plot the difference between simulated and emulated anisotropies, in units of  $\sigma$ , is presented. (a) Standard power-spectrum, for which simulated and emulated statistics are consistent. (b) Bispectrum, with a flattened triangle configuration, for which both simulation and emulation are statistically indistinguishable. (c) Histogram of pixel intensities, which are again extremely consistent. (d-f) These sub-figures display the three Minkowski functionals which are, from left to right, sensitive to the area, boundary, and Euler characteristic respectively. For both  $V_0$  and  $V_2$  simulation and emulation are highly consistent. However, a  $\sim 2\sigma$  discrepancy can be seen for  $V_1$  around  $\delta T/T/G\mu \approx 0$ . This exaggerated peak around 0 is likely due to low-intensity oscillations introduced from the extended support of bump steerable wavelets adopted in the PyWPH package. This effect could be mitigated by the use of alternative wavelets that are better localized in the spatial domain (which is beyond the scope of the current work). Nevertheless these summary statistics are overall in very good agreement.

use of synthesised observations for scientific inference, we adopt the method outlined in Section 3. We validate these synthetic cosmic string induced CMB anisotropies on a range of popular summary statistics that are sensitive to both Gaussian and non-Gaussian information content. Specifically we consider the power spectrum (Lizarraga et al. 2014a,b, 2016; Charnock et al. 2016), squeezed bispectrum (Planck Collaboration XXV 2014; Regan & Hindmarsh 2015), Minkowski functionals (Gott et al. 1990), and higher order statistical moments.

Looking to Figure 3, the power spectrum (Figure 3a), bispectrum with flattened triangle configuration  $B(k, k, k/2)$  (Figure 3b), and the distribution of pixel intensities (Figure 3c) are matched to well within  $1\sigma$  (grey region). The variance of these statistics accurately mirrors those computed on simulations indicating a similar degree of variability, which is encouraging.

The Minkowski functionals (Mecke et al. 1993) of a  $d$ -dimensional space are a set of  $d + 1$  functions that describe the morphological features of random fields. For 2-dimensional cosmic string maps  $d = 2$  and hence there exist three Minkowski functionals  $V_{0,1,2}$  which are sensitive to the area, boundary, and Euler characteristic of the excursion set respectively (an excursion set is simply the sub-set of pixels which are above some threshold magnitude). Looking again to Figure 3, we can see that  $V_0$  is recovered near perfectly (Figure 3d) and  $V_2$  is recovered to  $\sim 1\sigma$  (Figure 3f), however  $V_1$  is accurate away from  $\delta T/T/G\mu \approx 0$  but exhibits a  $\sim 2\sigma$  difference for  $\delta T/T/G\mu \approx 0$  (Figure 3e). Given that bump steerable wavelets do not have compact support in pixel-space (Allys et al. 2020), which can induce low-intensity extended oscillations, it is unsurprising that the error in  $V_1$  is largest around  $\delta T/T/G\mu \approx 0$ . An alternative family of wavelets could be considered with more compact support or, as mentioned in Section 3.1, total variation regularisation could be imposed to induce an inductive bias against such low-intensity oscillations. In fact, precisely such wavelet dictionaries have been developed on the sphere (Baldi et al. 2009; McEwen et al. 2018), however we leave exploration in this direction to future work.

Finally, in Figure 4 we consider a histogram of recovered skewness and kurtosis. It should be noted that the kurtosis in particular can be difficult to match, due to a high sensitivity to the tails of a distribution, which are often difficult to capture sufficiently (see e.g. Feeney et al. 2014). Nevertheless, we capture the distribution of both the skewness and kurtosis well.

In summary, although we find a moderate discrepancy for one statistic (the second Minkowski functional) around a single threshold (which could likely be mitigated in future by adopting different wavelets in the WPH representation, or subsequent evolutions thereof), all other statistics are excellently matched, both in terms of bias and relative variability.

## 5. CONCLUSIONS

In this article we consider generative modelling, highlighting the differences between its application to natural images and for physics. In contrast to typical use-cases for natural images, in physics it is important to not only generate realistic emulations but to also faithfully

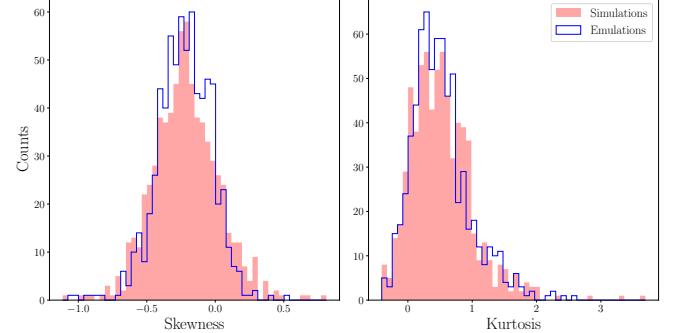


FIG. 4.— Histograms of the skewness and kurtosis respectively, generated from 700 instances of simulated and emulated Nambu-Goto cosmic string induced CMB anisotropies. We find an excellent agreement between the sets of emulated and simulated cosmic string induced CMB anisotropies.

trace the underlying probability distribution of fields. We ground this discussion within the context of cosmic string induced CMB anisotropies, which are structurally complex and highly computationally expensive to simulate. For scientific applications, generative models must not only generate realistic observables, but also ensure these synthetic observables are correctly distributed; a qualification which is often overlooked.

Leveraging the recently developed wavelet phase harmonics as a compressed latent representation, we present a method by which cosmic string induced anisotropies may accurately be synthesised at high-resolutions in under a minute. For context, flat-sky string simulations typically take more than a day to evolve, and full-sky simulations take in excess of 800,000 CPU hours. Importantly, our method requires significantly less data, which is a fundamental barrier for the application of many generative modelling techniques to cosmology. Our synthetic observations are statistically commensurate with those from simulated observations. In the spirit of reproducibility and accessibility our code has been made publicly available [Q](#).

Throughout, we consider the case where strings are generated from a Nambu-Goto action, however in principle the techniques we develop may equally be applied to other string models. For example, one may also emulate anisotropies induced by more complex scenarios such as cosmic superstring networks (e.g. Urrestilla & Vilenkin 2008). To accommodate fields with increased complexity, more expressive third generation scattering representations are likely to be useful (e.g. Cheng et al. 2023).

Although this work highlights the exciting potential for fast emulation of cosmic string induced CMB anisotropies, it is currently limited to the flat-sky. For wide-field observations (e.g. Planck) the sky curvature inevitably becomes non-negligible, hence the extension of these generative modelling techniques to the sphere is necessary. First generation wavelet scattering techniques on the sphere were developed in previous work (McEwen et al. 2022). In ongoing work we are developing accelerated and automatically differentiable spherical harmonic (Price et. al. 2023a in prep), wavelet transforms (Price et. al. 2023b in prep) and third generation spherical scattering covariances (Mousset et. al. 2023 in prep). Note that such third generation scattering covari-

ances have already shown much promise over flat spaces (Cheng et al. 2023). We are also exploring the fusion of these emulation techniques with simulation based inference, for application to many open areas of astrophysics.

#### ACKNOWLEDGEMENTS

The authors would like to thank Dipak Munshi, Tom Kitching, and Luke Pratley for discussion in the early stages of this work. Furthermore we would like to thank Erwan Allys for advice on wavelet phase harmonics, and both Christophe Ringeval and François Bouchet for providing the Nambu-Goto simulations, which are featured throughout this article. The cosmic string simulations have been performed thanks to computing support provided by the Institut du Développement des Ressources en Informatique Scientifique and the Planck-HFI processing center at the Institut d’Astrophysique de Paris. This work used computing equipment funded by the Research Capital Investment Fund (RCIF) provided by UKRI, and partially funded by the UCL Cosmoparticle Initiative. MAP and JDM are supported by EPSRC (grant num-

ber EP/W007673/1). MM and AM are supported by the STFC UCL Centre for Doctoral Training in Data Intensive Science (grant number ST/P006736/1). ASM is supported by the MSSL STFC Consolidated Grant (grant number ST/W001136/1) and the Leverhulme Trust.

#### CONTRIBUTION STATEMENT

Author contributions are specified below, following the Contributor Roles Taxonomy (CRediT).

MAP: Conceptualisation, Methodology, Software, Validation, Investigation, Supervision, Writing (Original Draft, Review & Editing);

MM: Methodology, Software, Validation, Investigation;

MMD: Validation, Data Curation, Visualisation;

ASM: Conceptualisation, Validation;

AM: Software, Data Curation;

JDM: Conceptualisation, Methodology, Validation, Supervision, Writing (Original Draft, Review & Editing).

#### REFERENCES

- Agarwal S., Abdalla F.B., Feldman H.A., Lahav O., Thomas S.A., 2012, Monthly Notices of the Royal Astronomical Society, 424, 2, 1409
- Albrecht A., Turok N., 1989, Phys. Rev., D40, 973
- Allen B., Shellard P., 1990, Phys. Rev. Lett., 64, 119
- Allys E., Levrier F., Zhang S., Colling C., Regaldo-Saint Blanchard B., Boulaenger F., Hennebelle P., Mallat S., 2019, Astronomy & Astrophysics, 629, A115
- Allys E., Marchand T., Cardoso J.F., Villaescusa-Navarro F., Ho S., Mallat S., 2020, Physical Review D, 102, 10, 103506
- Amari S.I., 2016, *Information geometry and its applications*, volume 194, Springer
- Amsel S., Berger J., Brandenberger R.H., 2008, J. Cosmol. Astropart. P., 4, 015, [0709.0982](#)
- Arjovsky M., Chintala S., Bottou L., 2017, in D. Precup, Y.W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 214–223
- Auld T., Bridges M., Hobson M., 2008, Monthly Notices of the Royal Astronomical Society, 387, 4, 1575
- Baldi P., Kerkyacharian G., Marinucci D., Picard D., 2009, Ann. Stat., 37 No.3, 1150, [arXiv:math/0606599](#)
- Bengio Y., Courville A., Vincent P., 2013, IEEE transactions on pattern analysis and machine intelligence, 35, 8, 1798
- Bennett D.P., Bouchet F.R., 1989, Phys. Rev. Lett., 63, 2776
- Bennett D.P., Bouchet F.R., 1990, Phys. Rev. D41, 2408
- Bevis H., Handley W., Fialkov A., de Lera Acedo E., Javid K., 2021, Monthly Notices of the Royal Astronomical Society, 508, 2, 2923
- Blanco-Pillado J.J., Olum K.D., Shlaer B., 2011, Phys.Rev., D83, 083514, [1101.5173](#)
- Bouchet F.R., Bennett D.P., Stebbins A., 1988, Nature, 335, 410
- Brandenberger R.H., 1994, International Journal of Modern Physics A, 9, 2117, [astro-ph/9310041](#)
- Brochard A., Zhang S., Mallat S., 2022, arXiv preprint arXiv:2203.07902
- Brown B.C., Caterini A.L., Ross B.L., Cresswell J.C., Loaiza-Ganem G., 2022, in *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*
- Bruna J., Mallat S., 2013, IEEE transactions on pattern analysis and machine intelligence, 35, 8, 1872
- Byrd R.H., Lu P., Nocedal J., Zhu C., 1995, SIAM Journal on scientific computing, 16, 5, 1190
- Chardin J., Uhrlrich G., Aubert D., Deparis N., Gillet N., Ocvirk P., Lewis J., 2019, Monthly Notices of the Royal Astronomical Society, 490, 1, 1055
- Charnock T., Avgoustidis A., Copeland E.J., Moss A., 2016, Phys. Rev. D, 93, 12, 123503, [1603.01275](#)
- Cheng S., Ménard B., 2021, Monthly Notices of the Royal Astronomical Society, 507, 1, 1012
- Cheng S., Morel R., Allys E., Ménard B., Mallat S., 2023, arXiv preprint arXiv:2306.17210
- Cheng S., Ting Y.S., Ménard B., Bruna J., 2020, Monthly Notices of the Royal Astronomical Society, 499, 4, 5902
- Chown R., et al., 2018, The Astrophysical Journal Supplement Series, 239, 1, 10
- Ciua R., Hernández O.F., 2017, J. Cosmol. Astropart. P., 2017, 08, 028, [arXiv:1706.04131](#)
- Ciua R., Hernández O.F., 2019, Mon. Not. Roy. Astron. Soc., 483, 4, 5179, [arXiv:1810.11889](#)
- Ciua R., Hernández O.F., 2020, Mon. Not. Roy. Astron. Soc., 492, 1, 1329, [arXiv:1911.06378](#)
- Ciua R., Hernández O.F., Wolman M., 2019, Mon. Not. Roy. Astron. Soc., 485, 1, 1377, [arXiv:1708.08878](#)
- Copeland E.J., Kibble T.W.B., 2009, Proceedings of the Royal Society of London Series A, 466, 623, [0911.1345](#)
- Crammer K., Brehmer J., Louppe G., 2020, Proceedings of the National Academy of Sciences, 117, 48, 30055
- Dai B., Wipf D., 2019, in *International Conference on Learning Representations*
- Danos R.J., Brandenberger R.H., 2010, International Journal of Modern Physics D, 19, 183, [0811.2004](#)
- De Bortoli V., Mathieu E., Hutchinson M., Thornton J., Teh Y.W., Doucet A., 2022, arXiv preprint arXiv:2202.02763
- Ducout A., Bouchet F.R., Colombi S., Pogosyan D., Prunet S., 2013, Mon. Not. Roy. Astron. Soc., 429, 2104, [1209.1223](#)
- Eickenberg M., et al., 2022, arXiv preprint arXiv:2204.07646
- Euclid Collaboration, et al., 2021, Monthly Notices of the Royal Astronomical Society, 505, 2, 2840, ISSN 0035-8711
- Feder R.M., Berger P., Stein G., 2020, Physical Review D, 102, 10, 103504
- Feeeney S.M., Marinucci D., McEwen J.D., Peiris H.V., Wandelt B., Cammarota V., 2014, J. Cosmol. Astropart. P., 2014, 1, 050, [arXiv:1308.0602](#)
- Fefferman C., Mitter S., Narayanan H., 2016, Journal of the American Mathematical Society, 29, 4, 983
- Fendt W.A., Wandelt B.D., 2007, The Astrophysical Journal, 654, 1, 2
- Fergusson J.R., Regan D.M., Shellard E.P.S., 2010, ArXiv e-prints, [1012.6039](#)
- Fraisse A.A., Ringeval C., Spergel D.N., Bouchet F.R., 2008, Phys. Rev., D78, 043535, [arXiv:0708.1162](#)
- Gott III J.R., Park C., Juszkiewicz R., Bies W.E., Bennett D.P., Bouchet F.R., Stebbins A., 1990, Astrophys. J., 352, 1
- Greig B., Ting Y.S., Kaurov A.A., 2022, Monthly Notices of the Royal Astronomical Society, 513, 2, 1719
- Hammond D.K., Wiaux Y., Vandergheynst P., 2009, Mon. Not. Roy. Astron. Soc., 398, 1317, [arXiv:0811.1267](#)
- Heitmann K., Higdon D., White M., Habib S., Williams B.J., Lawrence E., Wagner C., 2009, The Astrophysical Journal, 705, 1, 156
- Hergt L., Amara A., Brandenberger R., Kacprzak T., Refregier A., 2017, J. Cosmol. Astropart. P., 2017, 06, 004, [arXiv:1608.00004](#)
- Hindmarsh M., 1994, Astrophys. J., 431, 534, [astro-ph/9307040](#)
- Hindmarsh M., Lizarraga J., Urrestilla J., Dáverio D., Kunz M., 2017, Physical Review D, 96, 2, 023525
- Hindmarsh M.B., Kibble T.W.B., 1995, Reports on Progress in Physics, 58, 477, [hep-ph/9411342](#)
- Hockney R.W., Eastwood J.W., 2021, *Computer simulation using particles*, crc Press
- Horvat C., Pfister J.P., 2021, arXiv preprint arXiv:2105.12152

- Jeffrey N., Boulanger F., Wandelt B.D., Regaldo-Saint Blanchard B., Ally E., Levrier F., 2022, Monthly Notices of the Royal Astronomical Society: Letters, 510, 1, L1
- Jimenez R., Verde L., Peiris H., Kosowsky A., 2004, Phys. Rev. D, 70, 023005
- Kibble T.W.B., 1976, Journal of Physics A Mathematical General, 9, 1387
- Kingma D.P., Welling M., 2013, arXiv preprint arXiv:1312.6114
- Landriau M., Shellard E., 2003, Phys.Rev., D67, 103512, [astro-ph/0208540](#)
- Landriau M., Shellard E., 2011, Phys.Rev., D83, 043516, [1004.2885](#)
- Lawrence E., Heitmann K., White M., Higdon D., Wagner C., Habib S., Williams B., 2010, The Astrophysical Journal, 713, 2, 1322
- Lin Y.H., Hassan S., Blanchard B.R.S., Eickenberg M., Modi C., 2022, arXiv preprint arXiv:2210.14273
- Lizarraga J., Urrestilla J., Daverio D., Hindmarsh M., Kunz M., 2016, ArXiv e-prints, [1609.03386](#)
- Lizarraga J., Urrestilla J., Daverio D., Hindmarsh M., Kunz M., Liddle A.R., 2014a, Phys. Rev. Lett., 112, 17, 171301, [1403.4924](#)
- Lizarraga J., Urrestilla J., Daverio D., Hindmarsh M., Kunz M., Liddle A.R., 2014b, Phys. Rev. D, 90, 10, 103504, [1408.4126](#)
- Lo A.S., Wright E.L., 2005, ArXiv, [astro-ph/0503120](#)
- Loaiza-Ganem G., Ross B.L., Cresswell J.C., Caterini A.L., 2022, Transactions on Machine Learning Research
- Louis T., et al., 2014, Journal of Cosmology and Astroparticle Physics, 2014, 07, 016
- Mallat S., 2012, Communications on Pure and Applied Mathematics, 65, 10, 1331, [arXiv:1101.2286](#)
- Mallat S., Zhang S., Rochette G., 2020, Information and Inference: A Journal of the IMA, 9, 3, 721
- Marchand T., Ozawa M., Biroli G., Mallat S., 2022, arXiv preprint arXiv:2207.04941
- McEwen J.D., Durastanti C., Wiaux Y., 2018, Applied Comput. Harm. Anal., 44, 1, 59, [arXiv:1509.06767](#)
- McEwen J.D., Feeney S.M., Peiris H.V., Wiaux Y., Ringeval C., Bouchet F.R., 2017, Mon. Not. Roy. Astron. Soc., 472, 4, 4081, [arXiv:1611.10347](#)
- McEwen J.D., Wallis C.G.R., Mavor-Parker A.N., 2022, in *International Conference on Learning Representations, in press*, [arXiv:2102.02828](#)
- Mecke K.R., Buchert T., Wagner H., 1993, arXiv preprint astro-ph/9312028
- Moore J.N., Shellard E.P.S., Martins C.J.A.P., 2002, Phys. Rev., D65, 2, 023503, [hep-ph/0107171](#)
- Movahed M.S., Javanmardi B., Sheth R.K., 2013, Mon. Not. Roy. Astron. Soc., 434, 3597, [1212.0964](#)
- Mustafa M., Bard D., Bhimji W., Lukic Z., Al-Rfou R., Kratochvil J.M., 2019, Computational Astrophysics and Cosmology, 6, 1, 1
- Nelson D., et al., 2019, Computational Astrophysics and Cosmology, 6, 1, 1
- Nielsen F., 2020, Entropy, 22, 10, 1100
- Pen U.L., Seljak U., Turok N., 1997, Phys. Rev. Lett., 79, 1611, [astro-ph/9704165](#)
- Perraudin N., Marcon S., Lucchi A., Kacprzak T., 2021, Frontiers in Artificial Intelligence, 4, 673062
- Piras D., Joachimi B., Villaescusa-Navarro F., 2023, Monthly Notices of the Royal Astronomical Society, 520, 1, 668
- Planck Collaboration XXV, 2014, Astron. & Astrophys., 571, A25, [arXiv:1303.5085](#)
- Pope P., Zhu C., Abdelkader A., Goldblum M., Goldstein T., 2021, arXiv preprint arXiv:2104.08894
- Portilla J., Simoncelli E.P., 2000, International journal of computer vision, 40, 1, 49
- Ramachandra N., Valogiannis G., Ishak M., Heitmann K., Collaboration L.D.E.S., et al., 2021, Physical Review D, 103, 12, 123525
- Regaldo-Saint Blanchard B., Ally E., Boulanger F., Levrier F., Jeffrey N., 2021, Astronomy & Astrophysics, 649, L18
- Regaldo-Saint Blanchard B., Levrier F., Ally E., Bellomi E., Boulanger F., 2020, arXiv preprint arXiv:2007.08242
- Regan D., Hindmarsh M., 2015, J. Cosmol. Astropart. P., 10, 030, [1508.02231](#)
- Ringeval C., Bouchet F.R., 2012, Phys. Rev. D., 86, 2, 023513, [1204.5041](#)
- Ringeval C., Sakellariadou M., Bouchet F., 2007, JCAP, 0702, 023, [astro-ph/0511646](#)
- Rodriguez A.C., Kacprzak T., Lucchi A., Amara A., Sgier R., Fluri J., Hofmann T., Réfrégier A., 2018, Computational Astrophysics and Cosmology, 5, 1, 1
- Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B., 2022, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695
- Sadegh Movahed M., Khosravi S., 2011, J. Cosmol. Astropart. P., 3, 012, [1011.2640](#)
- Saydjari A.K., Portillo S.K., Slepian Z., Kahraman S., Burkhart B., Finkbeiner D.P., 2021, The Astrophysical Journal, 910, 2, 122
- Song Y., Ermon S., 2019, Advances in Neural Information Processing Systems, 32
- Song Y., Sohl-Dickstein J., Kingma D.P., Kumar A., Ermon S., Poole B., 2020, arXiv preprint arXiv:2011.13456
- Spurio Mancini A., Docherty M.M., Price M.A., McEwen J.D., 2022, Submitted to RASTI, [arXiv:2207.04037](#)
- Spurio Mancini A., Piras D., Alsing J., Joachimi B., Hobson M.P., 2022, Monthly Notices of the Royal Astronomical Society, 511, 2, 1771, ISSN 0035-8711
- Starck J.L., Aghanim N., Forni O., 2004, Astron. & Astrophys., 416, 9, [astro-ph/0311577](#)
- Stewart A., Brandenberger R., 2009, J. Cosmol. Astropart. P., 2, 009, [0809.0865](#)
- Torki M., Hajizadeh H., Farhang M., Vafaei Sadr A., Movahed S., 2022, Mon. Not. Roy. Astron. Soc., 509, 2, 2169, [arXiv:2106.00059](#)
- Urrestilla J., Vilenkin A., 2008, Journal of High Energy Physics, 2008, 02, 037
- Vafaei Sadr A., Farhang M., Movahed S., Bassett B., Kunz M., 2018, Mon. Not. Roy. Astron. Soc., 478, 1, 1132, [arXiv:1801.04140](#)
- Valogiannis G., Dvorkin C., 2022, Physical Review D, 105, 10, 103534
- Vilenkin A., Shellard E.P.S., 1994, *Cosmic strings and other topological defects*, Cambridge monographs on mathematical physics, Cambridge Univ. Press, Cambridge
- Villaescusa-Navarro F., et al., 2020, The Astrophysical Journal Supplement Series, 250, 1, 2
- Vincent G., Antunes N.D., Hindmarsh M., 1998, Phys. Rev. Lett., 80, 2277, [hep-ph/9708427](#)
- Wiaux Y., Puy G., Vandergheynst P., 2010, Mon. Not. Roy. Astron. Soc., 402, 4, 2626, [arXiv:0908.4179](#)
- Zhang S., Mallat S., 2021, Applied and Computational Harmonic Analysis, 53, 199

provides fast and easy peer review for new papers in the **astro-ph** section of the arXiv, making the reviewing process simpler for authors and referees alike. Learn more at <http://astro.theoj.org>.

This paper was built using the Open Journal of Astrophysics L<sup>A</sup>T<sub>E</sub>X template. The OJA is a journal which