Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Received: February 12, 2025
Revised: July 25, 2025
Accepted: August 14, 2025
Published: September 19, 2025

Cosmology with persistent homology: parameter inference via machine learning

Juan Calles[a,1] **Jacky H.T. Yip**[b,1] **Gabriella Contardo**[c,d] **Jorge Noreña**[e] **Adam Rouhiainen**[b] **and Gary Shiu**[b]

[a] *Instituto de Física y Astronomía, Universidad de Valparaíso, Avda. Gran Bretaña 1111, Valparaíso, Chile*
[b] *Department of Physics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.*
[c] *Center for Astrophysics and Cosmology, University of Nova Gorica, Ajdovščina I-5270, Slovenia*
[d] *Theoretical and Scientific Data Science, Scuola Internazionale Superiore di Studi Avanzati, Trieste 34136, Italy*
[e] *Instituto de Física, Pontificia Universidad Católica de Valparaíso, Casilla 4950, Valparaíso, Chile*

E-mail: juan.calles@uv.cl, hyip2@wisc.edu, gcontardo@ung.si, jorge.norena@pucv.cl, rouhiainen@wisc.edu, shiu@physics.wisc.edu

Abstract: Building upon previous work [1], we investigate the constraining power of persistent homology on cosmological parameters and primordial non-Gaussianity in a likelihood-free inference pipeline utilizing machine learning. We evaluate the ability of Persistence Images (PIs) to infer parameters, comparing them to the combined Power Spectrum and Bispectrum (PS/BS). We also compare two classes of models: neural-based and tree-based. PIs consistently lead to better predictions compared to the combined PS/BS for parameters that can be constrained, i.e., for $\{\Omega_{\rm m}, \sigma_8, n_{\rm s}, f_{\rm NL}^{\rm loc}\}$. PIs perform particularly well for $f_{\rm NL}^{\rm loc}$, highlighting the potential of persistent homology for constraining primordial non-Gaussianity. Our results indicate that combining PIs with PS/BS provides only marginal gains, indicating that the PS/BS contains little additional or complementary information to the PIs. Finally, we provide a visualization of the most important topological features for $f_{\rm NL}^{\rm loc}$ and for $\Omega_{\rm m}$. This reveals that clusters and voids (0-cycles and 2-cycles) are most informative for $\Omega_{\rm m}$, while $f_{\rm NL}^{\rm loc}$ is additionally informed by filaments (1-cycles).

Keywords: cosmological parameters from LSS, Machine learning

ArXiv ePrint: 2412.15405

[1] Equal contribution.

© 2025 The Author(s). Published by IOP Publishing Ltd on behalf of Sissa Medialab. Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

https://doi.org/10.1088/1475-7516/2025/09/064

JCAP09(2025)064

Contents

JCAP09(2025)064

1 Introduction

The large-scale structure (LSS) of the universe, traced by the distribution of galaxies, halos, filaments, and voids, encodes invaluable information about the underlying cosmological model, initial conditions, and the physical processes driving cosmic evolution. Cosmological parameter inference, which aims to extract this information, is central to advancing our understanding of the universe. Upcoming observational surveys such as SPHEREx [2], Euclid [3], LSST [4] promise to significantly enhance our ability to probe the LSS at groundbreaking precision. However, fully leveraging the wealth of information contained in the LSS remains a major challenge, particularly in the nonlinear regime where gravitational collapse gives rise to intricate, strongly non-Gaussian features.

Traditional summary statistics, such as the two-point correlation function and its Fourier transform, the power spectrum, have long been the conventional tools for analyzing the

LSS. These low-order statistics effectively capture the primary properties of the LSS, such as clustering and density fluctuations, and have been widely used to constrain parameters like the matter density $\Omega_{\rm m}$ and the amplitude of matter fluctuations σ_8 [5–14]. Higher-order statistics, such as the bispectrum, extend this framework by probing non-Gaussian features of the matter distribution [15–20]. However, these methods often struggle to capture the full complexity of the nonlinear, small-scale regime. Moreover, the question of which summary statistics are most effective for analyzing non-Gaussian fields remains unresolved, as different approaches may capture complementary aspects of the underlying physics.

This effort is motivated by the goal of uncovering the fundamental physical processes shaping the universe. Of particular interest for primordial cosmology is the study of the deviation of initial metric fluctuations from a Gaussian distribution, i.e. primordial non-Gaussianity (PNG). However, gravitational evolution also sources non-Gaussianities, and local phenomena, such as galaxy bias, can produce signals that are difficult to distinguish from those originating in the early universe [21]. This overlap creates significant challenges for traditional summary statistics, which are often unable to separate late-time effects from signatures of fundamental physics [22–24]. Thus, novel methodologies are needed to disentangle these contributions and reliably identify the unique fingerprints of PNG, particularly on nonlinear scales where the interplay between these processes becomes most complex.

Recognizing this limitation, a range of alternative summary statistics have been proposed to extend the reach of LSS analyses into the nonlinear regime. Among them [25–30], skew-spectra [31, 32], Wavelet Scattering Transform [33–36], one-point PDFs [37–39], Void Abundance [40–42], k-nearest neighbors [43, 44], Minkowski functionals [45–47]. This is but a small sample of the many methods being developed in this very active area of research. Additionally, field-level inference, which proposes to perform inference directly on the entire density field bypassing the need for summary statistics, presents a promising approach [48–52]. By leveraging the full information contained in the data, field-level inference methods theoretically offer the greatest potential for maximizing the extraction of cosmological parameters. However, they require heavy machine learning machinery combined with extensive simulations. Such methods often yield results that are difficult to interpret and validate, particularly in understanding which field features are responsible for constraining the parameters of interest. In addition, they strongly rely on building reliable and "accurate" simulations, which might be a problematic bottleneck, especially when integrating astrophysical processes. These challenges are compounded by the significant computational demands of building such simulations. Another avenue is using forward modelling of the distribution of galaxies using the effective field theory of LSS [53–56], but this is still limited to relatively large scales.

This highlights the need to develop efficient and interpretable summary statistics that balance the high-information content of field-level analysis with the computational efficiency of traditional methods. In this context, a technique from topological data analysis (TDA), persistent homology, has emerged as a powerful tool for studying complex data structures. It is particularly well-suited for cosmology, where the LSS is organized into a hierarchical web of halos, filaments, and voids [57]. By tracking the "birth" and "death" of topological features —such as clusters, loops, and cavities— across scales, persistent homology provides a unique, multi-scale description of the universe.

JCAP09(2025)064

The outputs of persistent homology, known as persistence diagrams (PDs), can be transformed into persistence images (PIs), which summarize topological features in a format amenable to machine learning and statistical inference. While PIs could be directly incorporated into Bayesian inference frameworks, practical challenges such as estimating high-dimensional covariance matrices and capturing higher-order correlations, which would require an intractable number of simulations, have motivated the use of machine learning techniques. Neural-networks based inference, in particular, offer a robust approach for extracting patterns from PIs and addressing these challenges [1]. PIs are interpretable because the topological features they represent correspond to physically meaningful structures in the cosmic web.

Persistent homology has been successfully applied to a variety of problems in cosmology. Among the first implementations are [58–60], with specific applications ranging from identifying primordial non-Gaussianity in N-body simulations [61–65], to analyzing weak lensing through cosmic shear simulations [66, 67] and constraining the effects of massive neutrinos on the matter field [68].

This paper is a continuation of [1], which pioneered combining computational topology with machine learning in the context of cosmology. Using a convolutional neural network model, we map persistence images to cosmological parameters, enabling the extraction of information beyond that captured by traditional summary statistics. For benchmarking, we compare the performance of PIs with constraints derived from the power spectrum and bispectrum combined, offering a comprehensive evaluation of their relative information content.

We extend prior efforts in several key ways. First, we employ high-fidelity simulations from larger volumes and three independent datasets, ensuring robust and realistic parameter recovery. Second, we expand the parameter space to include both standard cosmological parameters and primordial non-Gaussianity amplitudes, providing a broader test of PIs' constraining power. In addition to predicting cosmological parameters, our models estimate associated uncertainties, offering a more complete characterization of the inference process. Finally, we enhance the interpretability of our results by employing feature-scoring methods, such as gradient-boosted trees, to identify the specific features of PIs that contribute most to parameter constraints.

This paper is organized as follows. Section 2 introduces the persistent homology framework and the construction of persistence images, emphasizing their relevance for LSS analysis. In section 3 we outline the dataset we use to compute our summary statistics built from the persistent homology pipeline. Additionally, we describe the building of the traditional power spectrum and bispectrum, computed from the halo catalogs used in this work. Section 4 describes the machine learning models used for parameter inference, including convolutional neural networks and gradient-boosted trees. Section 5 presents the main results, comparing the performance of PIs with traditional summary statistics in constraining cosmological parameters. Within this section, we also examine the sensitivity of PIs to individual cosmological parameters using feature attribution methods. Finally, section 6 discusses the broader implications of our findings and outlines future directions, such as integrating persistent homology with simulation-based inference frameworks and applying these methods to galaxy survey data.

JCAP09(2025)064

2 Topological data analysis and persistent homology

Topological data analysis (TDA) is a collection of methods originating in algebraic topology, a branch of mathematics founded by Poincaré that studies the shape of data through its topological features. Central to TDA is the notion of homology, which captures topological features such as connected components, loops, and voids in a given space.

In particular, persistent homology is a tool that extracts and quantifies the persistence of these features, enabling the identification of robust, scale-independent topological signals in noisy or high-dimensional data. We refer the reader to [69–71] for comprehensive introductions to foundational developments in TDA.

Persistent homology can be applied to either discrete point sets or continuous fields, provided a notion of filtration is defined. Conceptually, a filtration is a family of nested sets parametrized by a filtration parameter (also called filtration time), denoted by ν. This parameter typically reflects a notion of proximity in the data space. For each value of this geometric parameter, there is a set in the filtration in which topological features can be identified. Hence, as the parameter varies, topological features come into existence, evolve, and eventually die at various parameter values. Persistence statistics can then be built from this history of topological evolution.

2.1 Persistent homology of the large-scale structure

In the context of cosmology, this tool can be used to capture the multi-scale topological characteristics of the large-scale structure, roughly as a collection of clusters, loops, and voids distributed hierarchically in scale. These cosmological structures arise within the distribution of dark matter, which can be traced by dark matter halos, the hosts of visible galaxies. Hence, one expects that persistent homology applied to the spatial distribution of these halos probes these structures. To build an intuition, one can roughly interpret the large-scale structure, when expressed in terms of homology groups, as follows: high-density halo clusters extending into walls and filaments (0-cycles), loop-like filamentary bridges connecting matter concentrations (1-cycles), and fully enclosed cavities forming the central regions of cosmic voids (2-cycles). These topological features are not necessarily one-to-one mappings with visually defined cosmic structures, but rather mathematically defined equivalence classes (cycles modulo boundaries) that reflect how matter is connected at multiple scales. Figure 1 offers a heuristic illustration of how these topological features may arise from halo distribution. The filtration values, which indicate the scales at which these features are born and die, encode geometric information that can be used to infer the underlying cosmological model.

Let us walk through the implementation.[1] Given a point cloud of halo positions in 3 dimensions, to define a filtration for it is to write down a set of rules that adds connections between the halos depending on the value of the non-negative[2] filtration parameter ν. Here, connections refer to the n-simplices where halos act as the vertices: the 1-, 2-, and 3-simplices, which are edges, triangular faces, and solid tetrahedra respectively. An intuitive description of a filtration is that ν represents spatial distance, and a simplex is added if its size is equal to, or smaller than, ν. Accordingly, more simplices are included as we dial ν, and this

[1] We refer the reader to [63] for a more comprehensive and gentle introduction to the same implementation.
[2] Hence the expression "filtration time".

JCAP09(2025)064

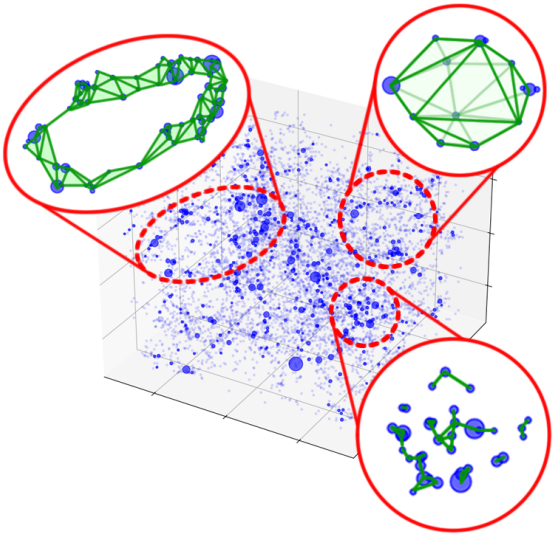

Figure 1. Cosmic structures are topological: Halo clusters (bottom right), filament loops (top left), and cosmic voids (top right) correspond to the 0-, 1-, and 2-cycles in topology, respectively.

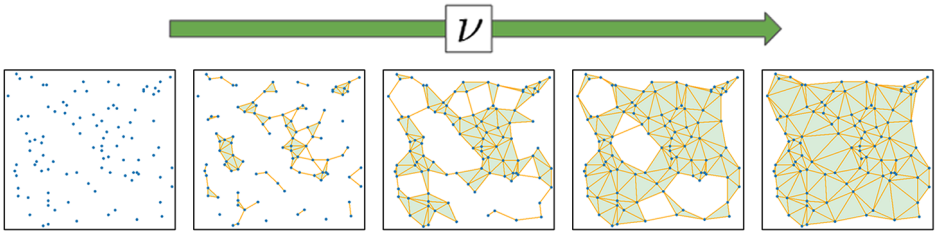

Figure 2. Example of a filtration for a point cloud in 2D, as a collection of simplicial complexes parametrized by the filtration parameter or time ν. As ν increases, n-simplices are subsequently added following a set of rules which can be flexibly designed. Topological features emerge and die throughout the filtration.

collection of simplices, known as a simplicial complex, becomes increasingly sophisticated in its connectedness, eventually leading to successive emergences of topological features. In other words, the filtration is the family of these simplicial complexes at varying ν's (figure 2).[3] For this paper, the n-simplices are taken from the Delaunay triangulation of the point cloud, and we adopt the α-DTM-ℓ filtration, which we explain in the next subsection. As a final note, formally known as a p-cycle, a topological feature is an equivalence class of boundaryless collections of p-simplices where each collection is not itself the boundary of any collection of $p+1$-simplices. Intuitively, these equivalent collections enclose the same $(p+1)$-dimensional "hole" that characterizes the topological feature.[4]

[3]The example shown is an α-filtration, which is what the α-DTM-ℓ filtration is based on.

[4]In plain terms, a 0-cycle is a collection of connected simplices, a 1-cycle is a loop, and a 2-cycle is an enclosed cavity.

JCAP09(2025)064

2.2 Lightning review of α-DTM-ℓ filtration

The α-DTM-ℓ filtration was first implemented on halos in [64]. It is based on the α-filtration (or α-shape) [72] defined by α-complexes [73], which is well-established and widely implemented by libraries such as `GUDHI` [74], which uses the `CGAL` library [75]. The α-filtration has previously been applied to the large-scale structures, see [58, 76, 77].

While the underlying simplicial structure remains based on Delaunay triangulations, the key feature of the α-DTM-ℓ filtration is that it uses the density-aware DTM (Distance-To-Measure) function to replace Euclidean distance in α-filtration. There is a parameter, the number of nearest neighbors used k,[5] adjustable for extracting topological information at different scales [63]. The set of rules defining the filtration is as follows:

1. We begin (at $\nu = 0$) with an empty simplicial complex, not even containing vertices (halos). A vertex x is added at $\nu_x = \mathrm{DTM}_x$, where

$$\mathrm{DTM}_x \equiv \sqrt{\frac{1}{k} \sum_{X_i \in \mathcal{N}_k(x)} \|x - X_i\|^2} \tag{2.1}$$

   is the Distance-To-Measure function which quantifies the sparsity around the halo that x corresponds to. Here $\mathcal{N}_k(x)$ is to x the set of k-nearest neighbors in the given halo point cloud, and $\|a - b\|$ is the Euclidean distance between vertex a and b. In other words, if the halo in question lies in a sparsely populated region, then DTM_x will be large, and x is added at a late filtration time.

2. An edge $\sigma_{x_1x_2}$ linking the vertices x_1 and x_2 is added if

$$d_{x_1x_2} \equiv \|x_1 - x_2\| \leq r_{x_1}(\nu) + r_{x_2}(\nu), \tag{2.2}$$

   where

$$r_x(\nu) \equiv \sqrt{\nu^2 - \mathrm{DTM}_x^2}. \tag{2.3}$$

   Alternatively, we can determine the time $\nu_{\sigma_{x_1x_2}}$ at which the edge $\sigma_{x_1x_2}$ is added by solving eqs. (2.2) and (2.3):

$$\nu_{\sigma_{x_1x_2}} = \sqrt{\frac{\left((\mathrm{DTM}_{x_1} + \mathrm{DTM}_{x_2})^2 + d_{x_1x_2}^2\right)\left((\mathrm{DTM}_{x_1} - \mathrm{DTM}_{x_2})^2 + d_{x_1x_2}^2\right)}{2d_{x_1x_2}}}. \tag{2.4}$$

3. Higher-dimensional simplices (2- and 3-simplices; triangles and tetrahedra) are added immediately when the necessary lower-dimensional faces (edges or triangules) are present. For example, if the edges $\sigma_{x_1x_2}$, $\sigma_{x_1x_3}$, and $\sigma_{x_2x_3}$ are added one after another, then the triangle $\sigma_{x_1x_2x_3}$ is also added at the same filtration time as $\sigma_{x_2x_3}$, i.e., $\nu_{\sigma_{x_1x_2x_3}} = \nu_{\sigma_{x_2x_3}}$.[6]

[5] Not to be confused with the wavenumber k associated with the power spectrum and bispectrum.

[6] Provided that the Delaunay triangulation of the given point cloud contains $\sigma_{x_1x_2x_3}$, which is not guaranteed for arbitrary combinations of x_1, x_2, and x_3.

JCAP09(2025)064

One can visualize Rule 2 as placing around each vertex x_i a growing sphere of radius $r_{x_i}(\nu)$, and the corresponding edge is added when two spheres touch or overlap. DTM_{x_i} impedes the growth of the sphere and delaying the inclusion of simplices involving x_i. As presented in [63], tuning k effectively regulates the average volume within which the algorithm explores for the densest regions to populate with simplices. Therefore, when a larger k is used, some regions are deemed no longer dense enough for early population, resulting in larger holes (i.e., longer-lived topological features) to emerge.[7] Intuitively, smaller values of k result in finer-scale, more locally sensitive filtrations, potentially capturing smaller features. In contrast, larger k values smooth out local density fluctuations, emphasizing more global topological features. In other words, we can vary k to change the scale at which topological information is extracted.

2.3 Persistence outputs: diagrams and images

With the filtration defined, we now have a simplicial complex that evolves with the filtration parameter, or the filtration time, ν. Topological features (holes) come into existence, persist, and are eventually trivialized (i.e., filled in by simplices). Using ν as the handle, we can quantify this topological history by tracking the life of every feature that has ever existed. Precisely, each feature is characterized by a tuple $(\nu_{\mathrm{birth}}, \nu_{\mathrm{persist}})$, where ν_{birth} is the filtration time at which the feature is formed, and $\nu_{\mathrm{persist}} = \nu_{\mathrm{death}} - \nu_{\mathrm{birth}}$ is the duration of its existence prior to trivialization. To sum up, the primary outputs of a persistent homology computation are lists of $(\nu_{\mathrm{birth}}, \nu_{\mathrm{persist}})$ pairs, and in our 3-dimensional application there are 3 such lists for the 0-, 1-, and 2-cycles. Each of these lists is often presented as *persistence diagrams* and *persistence images* (figure 3):

Persistence diagrams. We plot all the cycles that ever existed in the filtration in the ν_{persist}-ν_{birth} plane, generating persistence diagrams. There are 3 persistence diagrams in our application, one for each of the 3 homological dimensions.

Persistence images. Fully specifying an arbitrary persistence diagram requires $2\times$(number of cycles) values. In most scenarios, one would wish to work with a data vector of a fixed size. To this end, the conventional way of vectorizing a persistence diagram is to "pixelate" it into a persistence image. Specifically, we bin in two dimensions the ν_{persist}-ν_{birth} plane, by assigning a smoothing kernel to each point in the persistence diagram. Then we sum up all kernel contributions in each bin.

Persistence images have proven useful in large-scale structures analysis, such as [60]. There are other vectorizations that condense the information in a persistence diagram even further. An example is the concatenation of histograms of ν_{birth} and ν_{death} values, which is used in [62, 63], where a Gaussian likelihood is assumed in Fisher forecast setups. In this work, we use neural network models for inference. As a result, we are less constrained and need to summarize the persistence diagrams only minimally, allowing for a controlled amount of information loss.

[7] Watch filtrations with different k's in action: https://youtu.be/_phgkiZmY0c.

JCAP09(2025)064

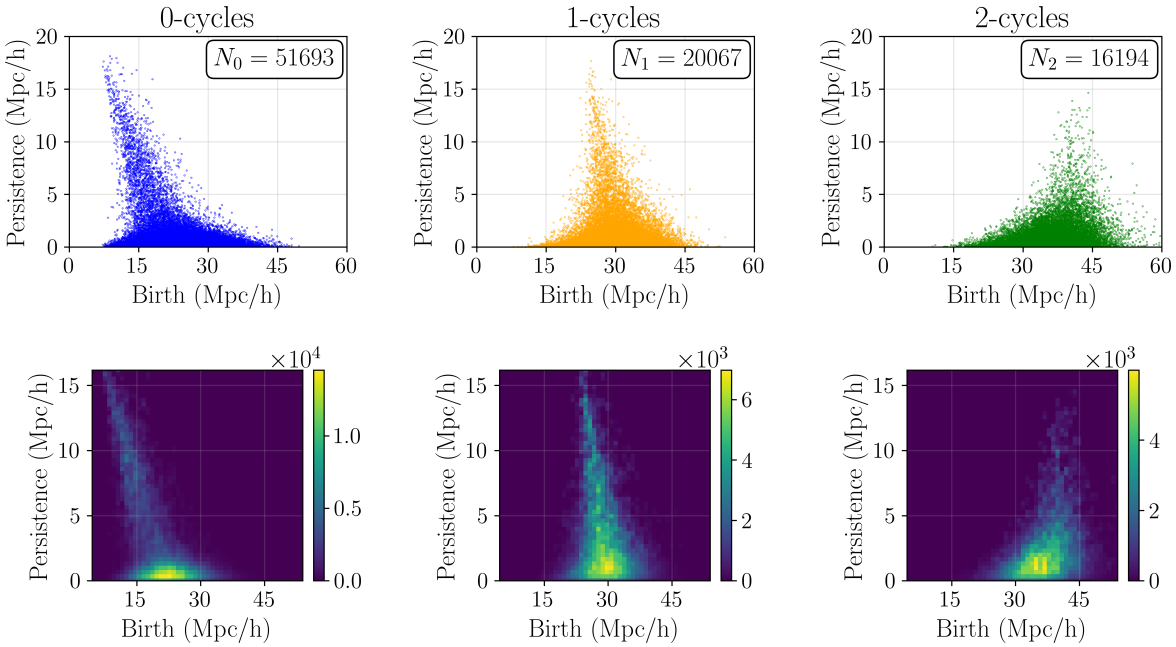

Figure 3. *Top Panel:* persistence diagrams of 0-, 1-, and 2-cycles from the filtration of a QUIJOTE halo catalog with $k = 15$ at the fiducial cosmology. N_p is the total number of p-cycles in each diagram, i.e., the total number of p-cycles that once existed in the filtration. *Bottom Panel:* corresponding persistence images from "pixelating" the diagrams.

3 Summary statistics from simulations

The entirety of our analysis is simulation-based. In this section, we provide a detailed description of how the persistence images, as well as the joint power spectrum and bispectrum statistic, are measured from the simulations.

3.1 Halo catalogs

We use the halo catalogs obtained from the N-body simulations in the QUIJOTE and QUIJOTE-PNG suites [78, 79] (hereafter collectively referred to as QUIJOTE). In each of these simulations, 512^3 dark matter particles are evolved using `GADGET-III` within a cosmological volume of $1\,(\mathrm{Gpc}/h)^3$, with initial conditions at redshift $z = 127$ generated by the public `2LPTIC` code. Dark matter halos are identified using the Friends-of-Friends (FoF) algorithm, and we utilize halo catalogs at $z = 0.5$, which are comparable to galaxies observed in surveys such as the BOSS CMASS sample. We also exclude halos that are composed of fewer than 50 particles, meaning that halos involved in our analysis have a minimum mass of $M_{\rm min} = 3.28 \times 10^{13} M_\odot$. We show in figure 4 a visual comparison of halo spatial structures and clustering morphologies driven by different cosmological parameters, using fixed initial conditions. For further details on the simulation suite, we refer the reader to existing studies of the QUIJOTE-PNG simulations [26, 80], which explicitly analyze the halo mass function with power spectrum and bispectrum statistics.

The training, validation, and testing datasets of summary statistics for our neural network inference models are measured from the Latin-hypercube (LH) subset of the QUIJOTE

JCAP09(2025)064

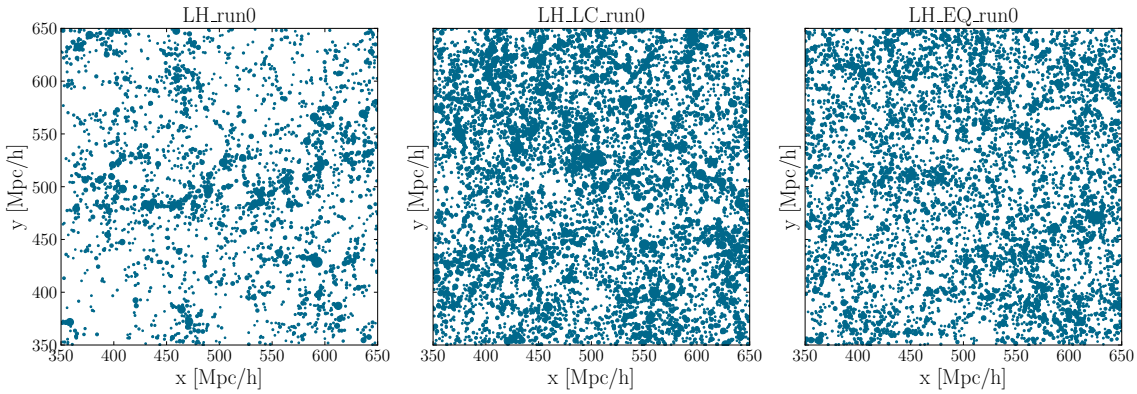

Figure 4. This figure shows a slice of thickness 250 Mpc/h along the z-direction, centered at 500 Mpc/h. Point sizes are proportional to halo mass. All panels share the same initial random seed but differ in cosmological parameters. The *left panel* corresponds to a realization from the LH dataset with $\Omega_{\rm m} = 0.1755$, $\Omega_{\rm b} = 0.06681$, $h = 0.7737$, $n_{\rm s} = 0.8849$, and $\sigma_8 = 0.6641$; the *middle panel* shows the same seed from the LH_$f_{\rm NL}^{\rm local}$ dataset, with $f_{\rm NL}^{\rm local} = -98.7$ and fiducial cosmological parameters. The *right panel* shows the LH_$f_{\rm NL}^{\rm equi}$ dataset, also with the same seed, adopting $f_{\rm NL}^{\rm equi} = 201$, $\Omega_{\rm m} = 0.3330$, $h = 0.5966$, $n_{\rm s} = 1.0726$, and $\sigma_8 = 0.7110$.

simulations.[8] Each simulation in an LH has a cosmology defined by cosmological parameters, including primordial non-Gaussianity amplitudes, which are sampled within specified ranges. We utilize 3 of the available LHs: LH, LH_$f_{\rm NL}^{\rm loc}$, and LH_$f_{\rm NL}^{\rm equi}$. For the fiducial parameter recovery tests, we employ the 15,000 fiducial realizations from the main suite. See table 1 for a summary.

Since observations are carried out in redshift space, prior to measurements, we convert the real-space positions of the halos to redshift-space positions. In the distant-observer approximation, the conversion is

$$\boldsymbol{x} \to \boldsymbol{x} + \frac{\boldsymbol{v}\cdot\hat{\boldsymbol{n}}}{a(z)H(z)}\hat{\boldsymbol{n}}, \tag{3.1}$$

where $\boldsymbol{x}$ is a halo's real-space position, $\boldsymbol{v}$ is the halo's peculiar velocity, $\hat{\boldsymbol{n}}$ is the line-of-sight, $a(z = 0.5) = (1 + 0.5)^{-1}$ is the scale factor, and the Hubble parameter $H(z) = 100\sqrt{\Omega_{\rm m}a^{-3} + \Omega_\Lambda a^{-3(1+w)}} = 100\sqrt{\Omega_{\rm m}a^{-3} + (1 - \Omega_{\rm m})}$ in (km/s)(h/Mpc) depends on $\Omega_{\rm m}$ of the catalog's cosmology. We process each catalog along the z-axis, i.e., taking $\hat{\boldsymbol{n}} = \hat{\boldsymbol{z}}$.

3.2 Persistence images

As described in section 2, we vary the nearest-neighbor parameter k in the α-DTM-ℓ filtration of our persistent homology pipeline. We choose $k \in \{1, 5, 15, 30, 60, 100\}$, which was shown in [1] to be an informative selection for Quijote's simulation volume and halo density; the persistence statistic derived from each value is expected to contain a reasonable amount of independent information. In summary, for a single halo catalog we have (3 homological dimensions) $\times$ (6 k values) $=$ 18 images.

We set the resolution of each persistence image to be 128×128 pixels, which provides a good trade-off between granularity and dimensionality given the size of our datasets. The

[8] https://quijote-simulations.readthedocs.io/en/latest/LH.html.

JCAP09(2025)064

Category	Number of realizations	$\Omega_{\rm m}$	$\Omega_{\rm b}$	h	$n_{\rm s}$	σ_8	$f_{\rm NL}^{\rm loc}$	$f_{\rm NL}^{\rm equi}$
Fiducial	15000	0.3175	0.049	0.6711	0.9624	0.834	0	0
LH	2000	$[0.1, 0.5]$	$[0.03, 0.07]$	$[0.5, 0.9]$	$[0.8, 1.2]$	$[0.6, 1.0]$	0	0
LH_$f_{\rm NL}^{\rm equi}$	1000	$[0.1, 0.5]$	0.049	$[0.5, 0.9]$	$[0.8, 1.2]$	$[0.6, 1.0]$	0	$[-600, 600]$
LH_$f_{\rm NL}^{\rm loc}$	1000	0.3175	0.049	0.6711	0.9624	0.834	$[-300, 300]$	0

Table 1. Simulation sets from Quijote and Quijote-PNG used in this work. The different LH (Latin-hypercube) realizations are used for the training and evaluation of the models, while the Fiducial simulations are used exclusively for the fiducial parameter recovery tests.

JCAP09(2025)064

bounds, i.e., minimum and maximum birth values, and maximum persistence values covered by an image are shared across images for the same homological dimension and value of k. We find the said bounds by surveying the LH set of simulations: for each realization we take the 1st and 99th percentile of all birth values and the 99th percentile of all persistence values, then we pool these values from all realizations and use the 1st/99th percentiles as the bounds. The first step removes outliers in each realization, while the second removes outliers within the simulation set. With this setup, we set for kernel density estimation the bandwidth parameter in the `KernelDensity` function in the `scikit-learn` library to be $5\times$ (persistence per pixel), where persistence per pixel is defined as the persistence bound divided by 128 (the 1D resolution).

3.3 Power spectrum and bispectrum

Following [63], we compute the redshift-space monopole and quadrupole components of the halo-halo power spectrum and the monopole of the bispectrum using the publicly available code PBI4.[9] The bin widths are defined as $\Delta k = 2k_f$, with the first bin centered at $2k_f$, where $k_f = 0.006\,h\mathrm{Mpc}^{-1}$ is the fundamental frequency of the simulation box. This binning procedure extends up to $k_{\rm max} = 0.3\,h\mathrm{Mpc}^{-1}$ beyond which shot noise dominates the signal, resulting in a total of 24 bins for each power spectrum component and 1522 triangular bins for the bispectrum monopole alone. To reduce aliasing effects, we implement the interlacing scheme of [81], combined with a fourth-order interpolation scheme, to compute the density contrast on a Fourier grid with 144 bins.

For both the power spectrum and the bispectrum monopole, we subtract a pure Poisson shot noise term, given by $1/\bar{n}$ for the power spectrum, and $1/\bar{n}\,(P(k_1) + P(k_2) + P(k_3)) + 1/\bar{n}^2$ for the bispectrum, where $\bar{n}$ is the halo density of the catalog.

4 Model architectures, training, and hyperparameter optimization

In this section, we present the different models used in our analysis, as well as their training and optimization strategies for predicting cosmological parameters from persistence images (PIs), power spectrum and bispectrum (PS/BS), and both combined.

[9] Available at: https://github.com/matteobiagetti/pbi4.

4.1 Convolutional neural networks for persistence images

Our generic CNN architecture combines two types of convolutional blocks. The first type consists of a Conv2D layer with a kernel size of 3, activated by LeakyReLU, followed by a max-pooling operation with a kernel size of 2. The second type is similar but removes the max-pooling operation. The first block is applied several times, followed by a couple of applications of the second block. The data is flattened after the convolutional layers to produce a one-dimensional vector and passed through a dropout layer before being fed into a fully connected output layer. This final layer predicts both the mean and standard deviation, as described below in section 4.5. The specific number of blocks and channels, as well as the dropout rates, weight decay, and learning rate were optimized through a hyperparameter search for individual datasets (i.e. LHs), as described in section 4.4.

We apply a 2D batch normalization layer directly to the input images for the specific case of the LH_$f_{\rm NL}^{\rm loc}$ dataset. In contrast, for the other Latin Hypercubes, we include the 2D batch normalization after every convolutional layer within each block, as we observed that this gives the best results.

4.2 Multi-Layer Perceptron for joint power spectrum and bispectrum

For the joint power spectrum and bispectrum statistics, we use a Multi-Layer Perceptron (MLP) model. The input consists of the concatenation of the total power spectrum and bispectrum statistics, denoted as $[P_0, P_2, B_0]$.

The MLP architecture applies a batch normalization for the input to ensure consistent scales across features. This is followed by several hidden layers, each composed of a linear layer with a LeakyReLU activation function and a dropout for regularization. The output layer is a linear layer that outputs the mean and the standard deviation for each parameter. We performed hyperparameter optimization on the learning rate, weight decay, number of neurons, layers and dropout rates.

While this general architecture was applied consistently to every Latin hypercube dataset, we explored various architectural modifications to assess their potential for improvement. For instance, we tested adding batch normalization after every hidden layer, as well as separating the power spectrum and bispectrum statistics into two distinct branches. In this latter approach, the power spectrum components $[P_0, P_2]$ and the bispectrum $[B_0]$ were processed independently using separate MLPs. The outputs of these branches were then concatenated into a single vector and passed through a final output layer for regression. However, none of these variations yielded significant improvements over the simpler design described above.

4.3 Joint PI-PS-BS architecture

To integrate persistence images (PIs) with power spectrum and bispectrum (PS/BS) statistics, we use a hybrid architecture that combines a convolutional neural network (CNN) for the PI data and a multi-layer perceptron (MLP) for the PS/BS data, referred to as Hybrid in the remainder of the paper.

The MLP branch processes the PS/BS input data using the previously described MLP architecture, excluding its output layer to focus on feature extraction. Similarly, the PI branch processes its input data through a CNN model, also excluding its output layer. The

JCAP09(2025)064

representations extracted from both branches are then flattened and concatenated into a single feature vector. This combined representation is passed through a final regression layer, composed of fully connected layers with LeakyReLU activation functions and dropout, which predict the means and standard deviations for each cosmological parameter.

To improve training efficiency and performance, HYBRID is initialized with pre-trained weights from the independently trained MLP and CNN models. For the output fully connected layers, we performed hyperparameter tuning on the number of layers, number of neurons, weight decay, and learning rate.

4.4 Training and hyperparameter optimization of neural network models

To optimize model performance across datasets, we performed hyperparameter search using Bayesian optimization through the `gp_minimize` function from the `scikit-optimize` library.[10] The search explored configurations such as the number of layers, dropout rates, and learning rates. Initial trials sampled hyperparameters randomly, followed by refined sampling through Bayesian optimization.

For training, we used the Adam optimizer without a learning rate scheduler. Weights were initialized using the Kaiming normal distribution for LeakyReLU activations, and biases were initialized to zero. We used early stopping with a patience of 100 epochs to prevent overfitting. For all models, we used a fixed batch size of 32 and minimized the validation loss. The search process started with 10 random trials to explore the hyperparameter space, followed by 40 additional trials using Bayesian optimization.

The hyperparameter space included the following configurations: for the MLP, the number of layers was defined within the range $[2, 6]$, and the hidden layer sizes were chosen between the values $[32, 64, 128, 256, 512]$. For CNN, the number of max-pooling Conv2D blocks was set in the range of $[3, 7]$, and the number of Conv2D blocks was defined within the range of $[1, 4]$. Furthermore, the number of channels for the Conv2D blocks was selected from $[8, 16, 32, 64, 128]$. For the HYBRID, the number of layers in the final output layer ranged from $[1, 8]$, with layer sizes selected from $[8, 16, 32, 64, 128, 256, 512]$. Dropout rates were chosen from $[0.3, 0.4, 0.5]$. The learning rate was sampled from a log-uniform distribution in the range $[10^{-4}, 10^{-2}]$, and weight decay was sampled log-uniformly from $[10^{-5}, 10^{-1}]$ for all models.

The data split was consistent across all models to ensure uniform conditions. For the LH dataset, the split consisted of 1600 training samples, 200 validation samples, and 200 test samples. For the LH_$f_{\rm NL}^{\rm loc}$ and LH_$f_{\rm NL}^{\rm equi}$ datasets, we used 600 training samples, with 200 each for validation and testing. The same random split was applied across all models for training, validation, and test sets. This ensures that each model was evaluated on identical cosmologies in the test set, allowing direct performance comparison across architectures (and thus summary statistics).

4.5 Likelihood-free parameter inference

For parameter inference, we do not compute the full posteriors of the parameters $p(\theta_i|\boldsymbol{X})$, where θ_i's represent the parameters we aim to recover. These are inferred from the summary statistic $\boldsymbol{X}$, which is derived from the halo catalog in the Latin hypercube dataset. Instead,

[10] scikit-optimize.github.io/stable/modules/generated/skopt.gp_minimize.html.

JCAP09(2025)064

our architecture predicts two values for each parameter: the mean $\hat{\mu}_i$ and the variance $\hat{\sigma}_i$ of the marginal posterior distribution. We use a custom loss function based on the logarithmic moment network methodology [82], which has recently gained attention for estimating model errors when a likelihood function is unavailable, as implemented in [26, 50, 83–92]. The specific form of the loss function is as follows:

$$\mathcal{L} = \sum_{i \in \text{params}} \log\left(\sum_{j \in \text{batch}} (\theta_{i,j} - \hat{\mu}_{i,j})^2\right) + \sum_{i \in \text{params}} \log\left(\sum_{j \in \text{batch}} \left((\theta_{i,j} - \hat{\mu}_{i,j})^2 - \hat{\sigma}_{i,j}^2\right)^2\right), \quad (4.1)$$

where $\theta_{i,j}$ is the true value of the i-th parameter for the j-th sample in the batch, and $\hat{\mu}_{i,j}$ and $\hat{\sigma}_{i,j}$ denote the predicted mean and variance, respectively. The logarithmic terms ensure that both the mean and variance are similarly weighted across all parameters, preventing the loss function from being dominated by the least accurately estimated parameters.

We also tested a simple Mean Squared Error (MSE) loss function, given by $\mathcal{L} = \frac{1}{N}\sum_{i \in \text{params}}(\theta_i - \hat{\mu}_i)^2$ to estimate the posterior means. In terms of architecture and results, the performance was comparable.

4.6 Gradient boosted trees

Although neural network-based methods are powerful and flexible methods to learn from complex data such as (natural) images or text, they can suffer from shortcomings on tabular data, where tree-based methods often outperform them in medium-sized datasets [93, 94]. Additionally, tree-based methods seem more robust to uninformative features. While we refer to our Persistent Homology-based summary statistic as a Persistence *Image*, it can actually be considered as tabular data as well: the location of a specific pixel has a meaning (birth scale and persistence). Power spectrum and bispectrum are also tabular data. Since we are in a relatively small dataset regime, where neural networks could overfit or require extreme regularization, we propose to use Gradient Boosted Trees (GBT) as an additional method to evaluate the predictive power of the summary statistics considered here and their robustness (or not) across types of methods (in our specific regime). We favor GBT over Random Forest as they have been shown to perform better for a comparable computational cost.

Another interesting advantage of decision tree-based models like `XGBoost` is their ability to evaluate feature importance, which can provide insights into how the model processes the features to predict. In section 5.4, we provide visualizations and analysis of the feature importance on some of our datasets.

We train GBT using the `XGBoost` library [95] for each data combination. When persistence images are included, the pixel values are flattened into a single array. Datasets are split as for the neural nets. We perform grid-search with cross-validation to find the hyperparameters (learning rate, maximum depth, number of estimators, subsampling, max child weight) that best fit the training data. We find that models with very shallow trees are generally preferred (with a maximum depth of five). The reason may be that this avoids overfitting. We minimize the RMSE to train these models and perform model selection on the validation. Since there is no off-the-shelf way to perform simulation-based inference techniques with boosted trees, we do not estimate the variance of each cosmological parameter. Therefore, these models are evaluated using only their RMSE and R^2 scores on the test set, and not the χ^2.

JCAP09(2025)064

It is valuable to note that the training of each model can be performed in a few minutes even in modest CPUs, which is significantly faster than neural-based approaches. However, they also require significantly more RAM.

5 Results

We now evaluate the performance of the different methods and summary statistics on the different datasets (LHs) considered. We note that each model was trained and validated on a separate subset of the dataset, and no significant difference between training and validation losses was observed, indicating that the networks generalized well without overfitting.

5.1 Evaluation metrics

To evaluate the performance of each model, we use several metrics to quantify the accuracy of the predicted means and variances for each cosmological parameter. These metrics include the coefficient of determination (R^2 score), the chi-squared statistic (χ^2), and the root-mean-square deviation (RMSE).

R^2 score. Measures the goodness of fit for the predicted mean, and therefore how well $\hat{\mu}_i$ tends to the true value θ_i for each parameter. It is given by

$$R^2(\theta, \hat{\mu}) = 1 - \frac{\sum_i (\theta_i - \hat{\mu}_i)^2}{\sum_i (\theta_i - \bar{\theta})^2}\,, \tag{5.1}$$

where the θ_i is the true parameter for the i-th sample, $\hat{\mu}_i$ is the predicted mean from the network model, and $\bar{\theta}$ is the average of the true parameter over the entire test set. The best score is reached at $R^2 = 1$, indicating a nearly perfect prediction, while $R^2 = 0$ means the prediction is no better than using the average value. Notice that R^2 can be negative meaning that the model performs worse than just predicting the mean.

χ^2 score. Evaluates how well the predicted variance $\hat{\sigma}$ matches the spread of the data. It is defined as

$$\chi^2(\theta, \hat{\mu}, \hat{\sigma}) = \frac{1}{N} \sum_i \frac{(\theta_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2}, \tag{5.2}$$

where N is the number of samples in the test set, θ_i is the true parameter value, $(\hat{\mu}_i, \hat{\sigma}_i)$ is the predicted mean and variance output by the model. A χ^2 score close to 1 indicates that the predicted uncertainties are well-calibrated. If χ^2 is significantly greater than 1, the model underestimates the uncertainties. On the other hand, if it is significantly less than 1, the model overestimates them.

RMSE. Quantifies the discrepancy between the predicted values generated by a model and the actual ground truth values, providing a comprehensive measure of the model's predictive performance relative to the true parameter values. A lower RMSE indicates a superior fit of the model to the data, thereby suggesting better predictive accuracy. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_i (\theta_i - \hat{\mu}_i)^2}. \tag{5.3}$$

JCAP09(2025)064

5.2 Varying cosmologies

We provide in figure 5 the predicted mean posteriors along with their associated standard deviation (std) for all 200 test points in the LH dataset, for the three summary statistics (PI, PSBS, PI-PSBS) and their respective neural architectures. The predicted mean values from the network are plotted on the y-axis and the true (target) values are plotted on the x-axis. The black diagonal line represents a perfect match between predicted and true values. The error bars illustrate standard deviation estimated by the model for each prediction. Additionally, the smaller lower panels show the residuals between the predicted posterior means and the ground truth values, with the black line indicating a perfect match between the two. We see that for parameters which are well estimated, that is $\Omega_{\rm m}$ and σ_8, the predicted values and variances give an accurate statistical description of the model performance. On the other hand, for paramters that are poorly estimated, that is $\Omega_{\rm b}$ and h the models are picking the mean among realizations, and though this is somewhat accounted for in the large error bars, there is a large bias visible in the residuals. For $n_{\rm s}$ the models capture some of the trend, assigning large error bars and a large bias. We also observe that none of these models have large outliers, but there is some slight but noticeable bias in the prediction of $\Omega_{\rm m}$ when using the PSBS data.

To quantify the performance of these models, we summarize the metrics — Root Mean Squared Error (RMSE), R^2, and χ^2 in the following tables: table 2 for the main Latin Hypercube (LH) dataset, table 3 and table 4 for the LH_$f_{\rm NL}^{\rm loc}$ dataset and LH_$f_{\rm NL}^{\rm equi}$ datasets.

Results for standard cosmological parameters.

- Persistence images performed best for $\{\Omega_{\rm m}, \sigma_8\}$. In numbers, using PIs with the CNN achieved $\rm R^2$ scores of $\{0.96, 0.99\}$, compared to the MLP's (using PS/BS) scores of $\{0.89, 0.93\}$. On the other hand, none of the statistics are sensitive to h and $\Omega_{\rm b}$, while both statistics achieve similar performance for $n_{\rm s}$. Looking at the χ^2 values obtained, both CNN and MLP standard deviations seem to be well calibrated.

- When combining both summary statistics in the Hybrid architecture, we find that the RMSE on each parameter is close to that obtained using the data that best constrains it. This indicates that the power spectrum and bispectrum do not contain complementary information or not contained in the persistence images, as combining the three does not lead to significant improvement.

- Boosted trees lead to poorer performance than the neural networks, especially when involving persistence images (either alone or combined). We find that though they are cheaper to train, they tend to overfit in this setting, which impacts their generalization. This can be corrected by constraining the model flexibility, however at the cost of performance. Perhaps surprisingly, the neural-based methods seem to provide a better trade-off in this regime.

- We claim that in our setup persistent homology in general does not probe scales beyond $2\pi/k_{\rm max} \approx 21\,{\rm Mpc}/h$,[11] the scale at which the power spectrum and bispectrum are

[11] A more detailed discussion can be found in [63], section 5.1.

JCAP09(2025)064

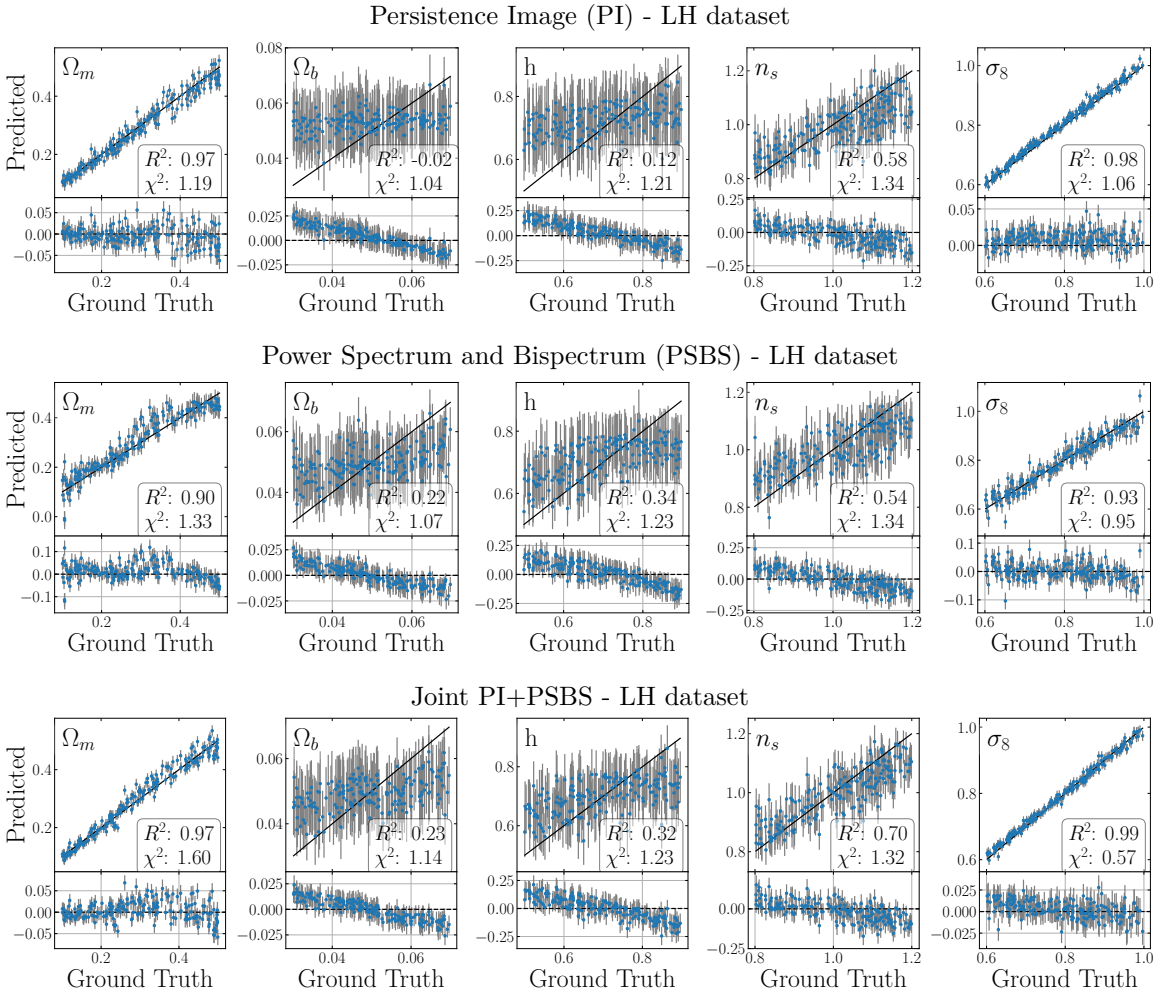

Figure 5. Model performance on predictions for all 200 test points using the best-performing models trained on the PI, PSBS, and PI-PSBS datasets. The black diagonal line represents a perfect match between predicted and true values. The error bars indicate the predicted standard deviation output by the network. The residuals are shown in the smaller lower panels. The dataset is the Quijote Latin Hypercube (LH) at redshift $z = 0.5$.

truncated. While it is difficult to define a scale for a topological feature, we can use the birth value, which for the 1- and 2-cycles is half of the length of the edge that triggers the formation of the feature. Figure 3 shows that most 1- and 2-cycles are born beyond $25\,\mathrm{Mpc}/h$, and a topological feature should be of a size larger than this edge. For the 0-cycles, the birth value corresponds inversely to the local halo number density, which we find to be within 20–$25\,\mathrm{Mpc}/h$, consistent with $\gtrsim 21\,\mathrm{Mpc}/h$, for the majority of the features. Hence, when persistent homology performs better than the power spectrum and bispectrum, we argue that the extra information extractable by the neural network model should not come from smaller-scale modes, but from the fact that persistent homology probes higher-order correlations given that each topological feature is typically formed from a set of many vertices.

JCAP09(2025)064

Data	Metric	$\Omega_{\rm m}$	$\Omega_{\rm b}$	h	$n_{\rm s}$	σ_8
PI (CNN)	RMSE	0.025 ± 0.002	0.011 ± 0.001	0.11 ± 0.001	0.074 ± 0.006	0.012 ± 0.001
	χ^2	1.31 ± 0.20	1.04 ± 0.09	1.12 ± 0.06	1.14 ± 0.19	0.77 ± 0.12
	R^2	0.96 ± 0.01	-0.064 ± 0.09	0.14 ± 0.02	0.59 ± 0.06	0.99 ± 0.01
PI (GBT)	RMSE	0.04 ± 0.01	0.01 ± 0.01	0.11 ± 0.03	0.09 ± 0.02	0.017 ± 0.003
	R^2	0.877 ± 0.002	-0.031 ± 0.0004	0.048 ± 0.003	0.387 ± 0.002	0.977 ± 0.001
$P+B$ (MLP)	RMSE	0.04 ± 0.001	0.01 ± 0.0001	0.095 ± 0.002	0.078 ± 0.002	0.029 ± 0.001
	χ^2	1.42 ± 0.14	1.02 ± 0.05	1.19 ± 0.09	1.26 ± 0.15	0.91 ± 0.14
	R^2	0.89 ± 0.01	0.24 ± 0.02	0.31 ± 0.03	0.55 ± 0.02	0.93 ± 0.01
$P+B$ (GBT)	RMSE	0.039 ± 0.009	0.01 ± 0.05	0.10 ± 0.04	0.08 ± 0.02	0.024 ± 0.002
	R^2	0.897 ± 0.001	0.162 ± 0.001	0.223 ± 0.004	0.453 ± 0.003	0.952 ± 0.001
Combined (HYBRID)	RMSE	**0.023 ± 0.002**	0.01 ± 0.001	0.096 ± 0.003	**0.068 ± 0.005**	**0.011 ± 0.002**
	χ^2	**1.47 ± 0.16**	1.01 ± 0.11	1.15 ± 0.13	**1.40 ± 0.24**	**0.80 ± 0.19**
	R^2	**0.97 ± 0.01**	0.21 ± 0.10	0.30 ± 0.05	**0.66 ± 0.05**	**0.99 ± 0.01**
Combined (GBT)	RMSE	0.040 ± 0.001	0.011 ± 0.001	0.110 ± 0.004	0.093 ± 0.003	0.017 ± 0.001
	R^2	0.893 ± 0.006	-0.024 ± 0.024	0.08 ± 0.06	0.36 ± 0.03	0.975 ± 0.002

Table 2. Performance metrics for CNN, MLP, HYBRID (on their respective summary statistics) and GBT (for the different statistics) for the LH dataset. Each value is the mean over 10 model initializations and the corresponding standard deviation. For boosted trees, the standard deviation comes from cross-validation with 4 different validation sets, keeping the training set size fixed.

PNG results.

- Persistence images consistently perform better than the combination of power spectrum and bispectrum for retrieving the $f_{\rm NL}^{\rm loc}$ parameter. The performance is significantly better when using GBT, unlike the previous results on the LH. This observation requires more investigation to properly understand its underlying causes. This could be due to a combination of the variance differences between the two datasets (one varying 5 parameters, 3 of which have an effective impact on the summary statistics considered here, the other varying only a single parameter), the (relatively) small size of each dataset, the nature of the data and the properties of the methods.

- All models struggled or failed to recover the $f_{\rm NL}^{\rm equi}$ value, and we therefore do not plot the Fiducial recovery. This seems to indicate that the summary statistics considered here are not sensitive enough to $f_{\rm NL}^{\rm equi}$ when marginalizing over the other parameters. Note that the $f_{\rm NL}^{\rm equi}$ inference case is much harder than the $f_{\rm NL}^{\rm loc}$ case. The LH_$f_{\rm NL}^{\rm equi}$ varies several cosmological parameters in addition to $f_{\rm NL}^{\rm equi}$, thus requiring to marginalize over those parameters. It would be interesting to investigate in future work whether integrating information about the other cosmological parameters (e.g., conditioning on the other parameters) helps in predicting $f_{\rm NL}^{\rm equi}$ with those summary statistics.

- It is also important to note that $f_{\rm NL}^{\rm equi}$ is defined to give a primordial bispectrum of the same size as $f_{\rm NL}^{\rm loc}$ at the equilateral configuration [96]. However, the local template is small for that configuration relative to other (more squeezed) configurations, so for

JCAP09(2025)064

Model	Metric	$f_{\rm NL}^{\rm loc}$
PI (CNN)	RMSE	47 ± 2
	χ^2	1.48 ± 0.17
	R^2	0.93 ± 0.01
PI (GBT)	RMSE	38.3 ± 1.2
	R^2	0.950 ± 0.03
$P+B$ (MLP)	RMSE	50 ± 2
	χ^2	2.40 ± 0.26
	R^2	0.92 ± 0.01
$P+B$ (GBT)	RMSE	48.8 ± 3.2
	R^2	0.920 ± 0.01
Combined (HYBRID)	RMSE	46 ± 2
	χ^2	1.58 ± 0.17
	R^2	0.93 ± 0.01
Combined (GBT)	RMSE	**37.6 ± 1.1**
	R^2	**0.952 ± 0.003**

Table 3. Performance metrics for LH_$f_{\rm NL}^{\rm loc}$ on the test set for the different summary statistics and models. For neural networks, each value is the mean over 10 model initializations, with the standard deviation reported. For boosted trees, the standard deviation comes from 4-fold cross-validation.

equal values of $f_{\rm NL}^{\rm loc}$ and $f_{\rm NL}^{\rm equi}$, local PNG has a larger physical effect. This effect is more pronounced for statistics that sum over many configurations, such as persistence images.

- The other results on the LH_$f_{\rm NL}^{\rm equi}$ dataset show a similar trend to the LH results regarding the informativeness of the persistence images. However, here we achieve best performance using the persistence images alone, with CNN, for the parameters that can be constrained, such as $\{\Omega_{\rm m}, \sigma_8, n_{\rm s}\}$. GBT consistently underperform compared to the neural-based methods. Combining the PIs with the PS/BS in this setup leads to worse results. One possible explanation is that this setup enters a regime where the high dimensionality of the data, coupled with the smaller dataset size (half the size of the LH dataset), prevents the models from training properly.

5.3 Fiducial parameter recovery

We now turn our attention to the fiducial dataset. We want to evaluate how well our models constrain parameters at the fiducial cosmology. The dataset contains 15,000 fiducial realizations never seen by the models, and from them we generate the same number of predictions for each best-performing model. The predictions are plotted in figures 6 and 7, for models trained on LH and LH_$f_{\rm NL}^{\rm loc}$ respectively (we do not plot for the models trained on LH_$f_{\rm NL}^{\rm equi}$ as they fail to infer $f_{\rm NL}^{\rm equi}$ completely), as contours marking the 68% and 95% confidence regions, as well as each parameter's 1D distribution. We report the following results:

JCAP09(2025)064

Model	Metric	$f_{\rm NL}^{\rm equi}$	$\Omega_{\rm m}$	h	$n_{\rm s}$	σ_8
PI (CNN)	RMSE	340 ± 2	**0.029 ± 0.005**	0.11 ± 0.002	**0.078 ± 0.005**	**0.018 ± 0.002**
	χ^2	1.02 ± 0.02	**1.02 ± 0.28**	1.17 ± 0.10	**1.19 ± 0.17**	**1.01 ± 0.23**
	R^2	0.005 ± 0.009	**0.94 ± 0.03**	0.22 ± 0.04	**0.55 ± 0.06**	**0.98 ± 0.01**
PI (GBT)	RMSE	338 ± 10	0.050 ± 0.003	0.109 ± 0.004	0.096 ± 0.007	0.023 ± 0.002
	R^2	-0.01 ± 0.03	0.81 ± 0.03	0.13 ± 0.07	0.30 ± 0.05	0.963 ± 0.005
$P+B$ (MLP)	RMSE	340 ± 0.3	0.04 ± 0.001	0.089 ± 0.002	0.084 ± 0.003	0.035 ± 0.002
	χ^2	0.96 ± 0.01	0.98 ± 0.07	1.36 ± 0.08	1.14 ± 0.07	1.11 ± 0.12
	R^2	0.002 ± 0.002	0.88 ± 0.01	0.42 ± 0.02	0.47 ± 0.04	0.91 ± 0.01
$P+B$ (GBT)	RMSE	333 ± 14	0.042 ± 0.001	0.104 ± 0.003	0.081 ± 0.006	0.028 ± 0.004
	R^2	0.02 ± 0.02	0.86 ± 0.01	0.21 ± 0.06	0.50 ± 0.06	0.94 ± 0.01
Combined (HYBRID)	RMSE	340 ± 0.3	0.04 ± 0.001	0.089 ± 0.002	0.084 ± 0.003	0.035 ± 0.002
	χ^2	0.96 ± 0.01	0.98 ± 0.07	1.36 ± 0.08	1.14 ± 0.07	1.11 ± 0.12
	R^2	0.002 ± 0.002	0.88 ± 0.01	0.42 ± 0.02	0.47 ± 0.04	0.91 ± 0.01
Combined (GBT)	RMSE	338 ± 10	0.046 ± 0.003	0.109 ± 0.004	0.092 ± 0.007	0.023 ± 0.006
	R^2	-0.01 ± 0.04	0.84 ± 0.02	0.13 ± 0.07	0.35 ± 0.06	0.962 ± 0.005

Table 4. Performance Metrics Comparison for LH_$f_{\rm NL}^{\rm equi}$ on the test set, for the different summary statistics and models. For neural networks, each value is the mean over 10 model initializations and the standard deviation over the mean. For boosted trees, the standard deviation comes from 4-fold cross-validation.

- Persistence images outperform power spectrum and bispectrum combined in both accuracy and precision for $\{\Omega_{\rm m}, n_{\rm s}, \sigma_8\}$. This is expected since the model achieves better R^2 values for those parameters.

- For $\{\Omega_{\rm b}, h\}$, predictions from the power spectrum and bispectrum combined are well centered on the fiducial value; however, this should not be overinterpreted, as the models have poor R^2 scores overall for these parameters. The model simply learns better at using the mean of the parameter values it trains on as the prediction, which happens to be the fiducial value.

- The HYBRID model generally does not improve on either of the other models that use a single type of statistics, except for $\{\Omega_{\rm m}, f_{\rm NL}^{\rm loc}\}$.

- The contours seem offset with respect to the fiducial values for parameters which are not well constrained. This does not necessarily mean that the model is biased since the contours do not include information about the estimated variance on the prediction.

5.4 Understanding persistence images with feature importance

Features importance is a potentially useful tool for understanding how machine learning models link input data to the outcomes. From tree-based models, it is possible to extract an importance score for each input feature, quantifying its contribution to the model's performance.

JCAP09(2025)064

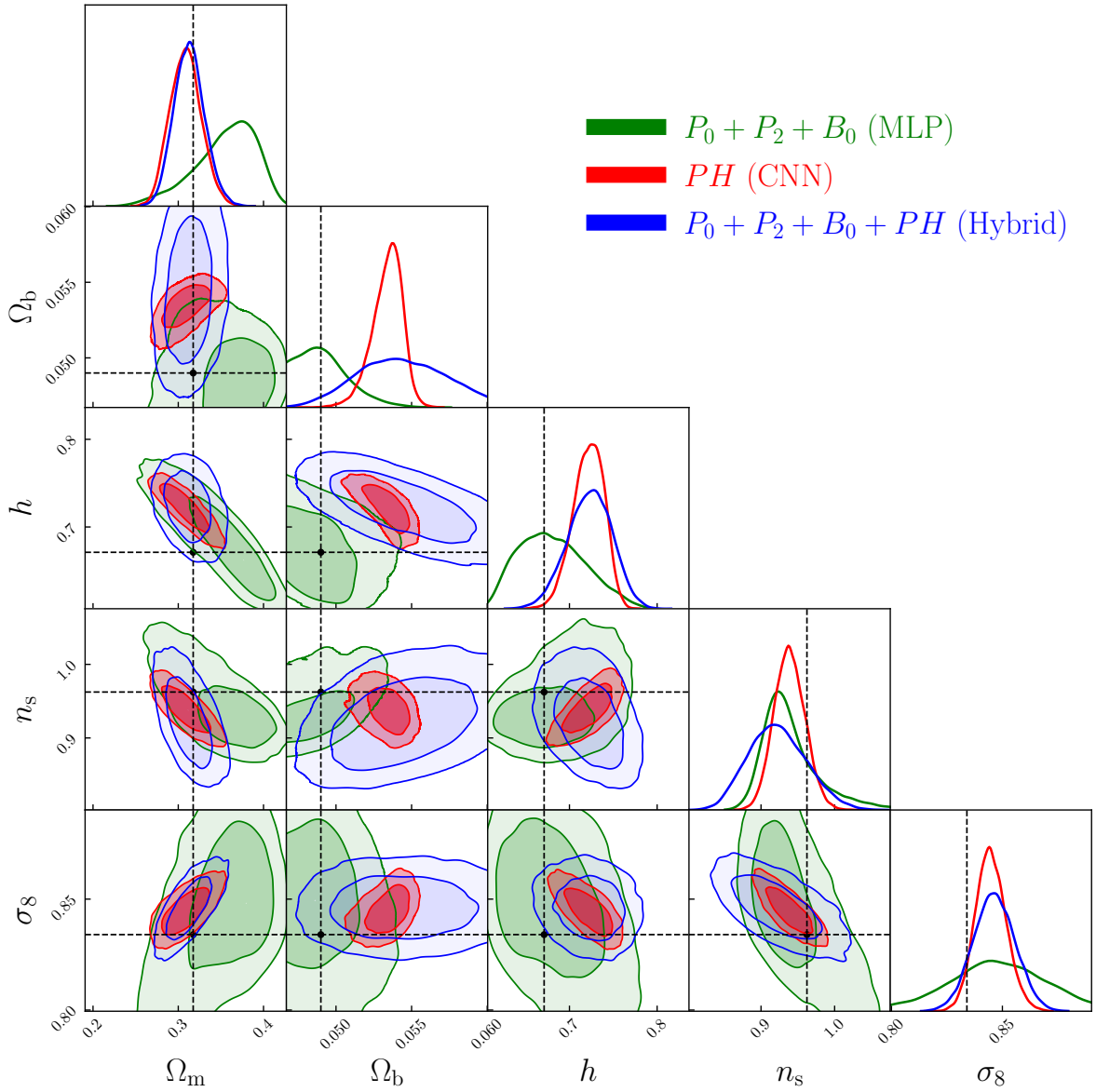

Figure 6. Predictions of cosmological parameters on the Fiducial dataset using the best-performing models trained on the Quijote Latin Hypercube (LH) at redshift $z = 0.5$. The contours indicate the regions containing 68% (1-σ) and 95% (2-σ) of the predictions. The crosshairs indicate the true fiducial values. The diagonal panels show the 1D distributions.

The feature importance map generated by the GBT model identifies the pixels with the greatest influence on the model's predictions. In `XGBoost`, the "weight" of a feature refers to the number of times that feature is selected for a split across all decision trees in the ensemble, thus essentially measuring how frequently a feature contributes to the decision-making process within the model. Each tree participating in the ensemble only uses a few features, such that only a few pixels are assigned a non-zero weight. This score is computed directly by the `XGBoost` package. In figure 8, we plot a map for each of the 3 input cycles and for each of the k — which determines the number of neighbors considered in the filtration, as described

JCAP09(2025)064

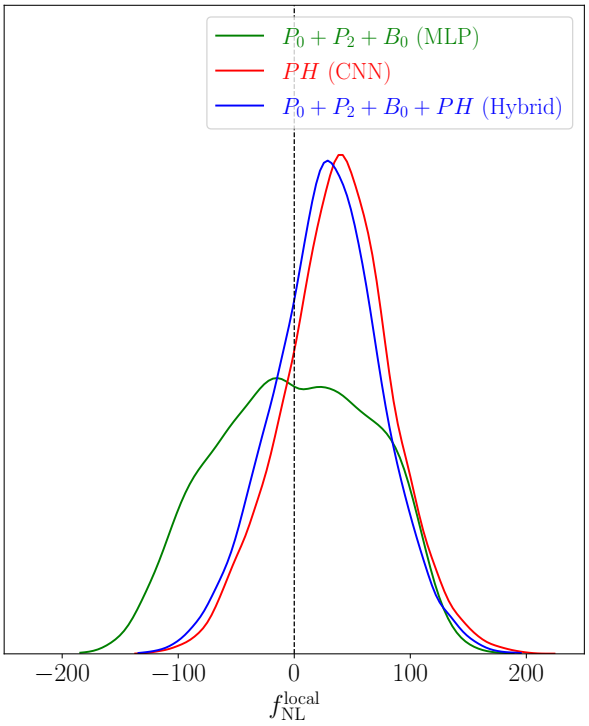

Figure 7. Distribution of $f_{\rm NL}^{\rm loc}$ predictions using the best-performing models trained on the Quijote-PNG Latin Hypercube (LH_$f_{\rm NL}^{\rm loc}$) at redshift $z = 0.5$. The dashed line marks the fiducial value $f_{\rm NL}^{\rm loc} = 0$.

earlier. For visualization purposes, the background shows the corresponding persistence image for a realization with $f_{\rm NL}^{\rm loc} = -288.3$. Crosses mark the positions of important features for predicting $f_{\rm NL}^{\rm loc}$, with the color indicating their importance score: cyan crosses correspond to non-null pixels with a weight below 1, whereas pink crosses represent pixels with a weight above 1. Figure 9 shows a similar plot for a model trained to predict $\Omega_{\rm m}$ only.

This visualization shows that regions associated with early-birth, low-persistence features in 0-cycles, as well as the borders of the persistence image, are critical for the model's predictions. For example, these areas exhibit the largest variations in the $f_{\rm NL}^{\rm loc}$ parameter (see figure 8 in [63]). These regions emphasize information tied to overdense regions, particularly those associated with early-forming structures. These features, identified early in the persistence pipeline, are often associated with "independent" massive halos. We see that the model trained on $f_{\rm NL}^{\rm loc}$ uses a relatively greater number of features from 1-cycles compared to $\Omega_{\rm m}$. Most of the information on $\Omega_{\rm m}$ seems to be contained in the 0- and 2-cycles (which roughly correspond to clusters and voids respectively).

This is a first attempt at understanding the persistence image summary statistic. It would be interesting to explore the robustness of these conclusions with a larger dataset.

6 Conclusions and outlook

This paper investigates the potential of persistence images (PIs), along with their combination with power spectrum and bispectrum (PS/BS) data, for cosmological parameter inference. We trained and compared these summary statistics using different types of machine learning

JCAP09(2025)064

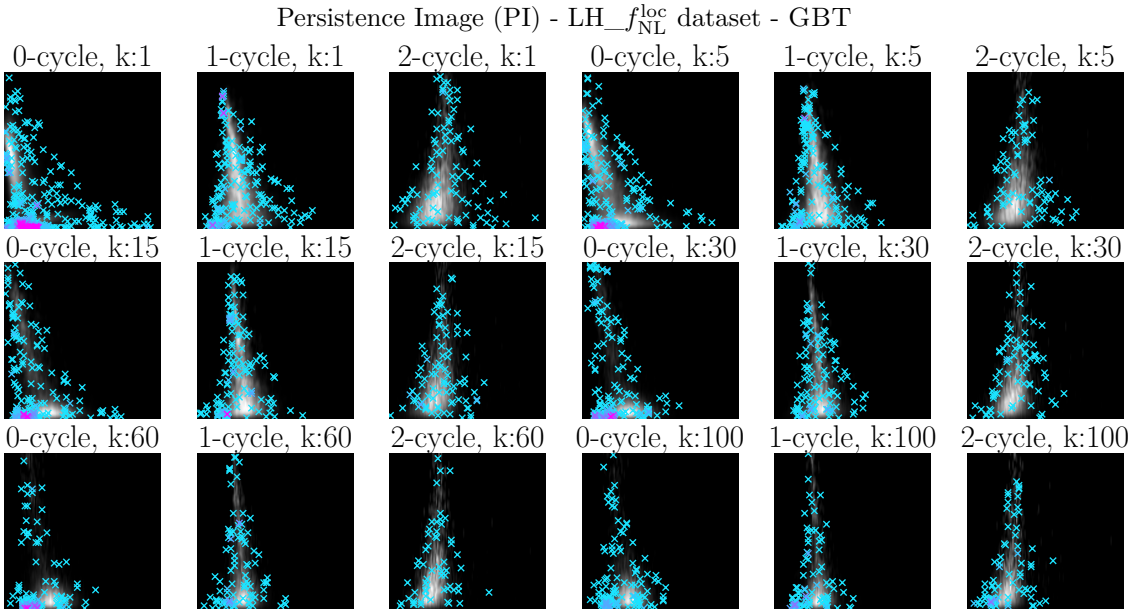

Figure 8. Feature importance map for the GBT model trained on the LH_$f_{\rm NL}^{\rm loc}$ dataset to predict $f_{\rm NL}^{\rm loc}$. Crosses indicate the positions of non-null features, with cyan marking pixels where the feature weight is non-null and below 1, and pink denoting weights above 1. The background shows the persistence image for a realization with $f_{\rm NL}^{\rm loc} = -288.3$, providing context for the highlighted features. Each image represents a different input channel, corresponding to all cycles with varying k-neighbors in the filtration.

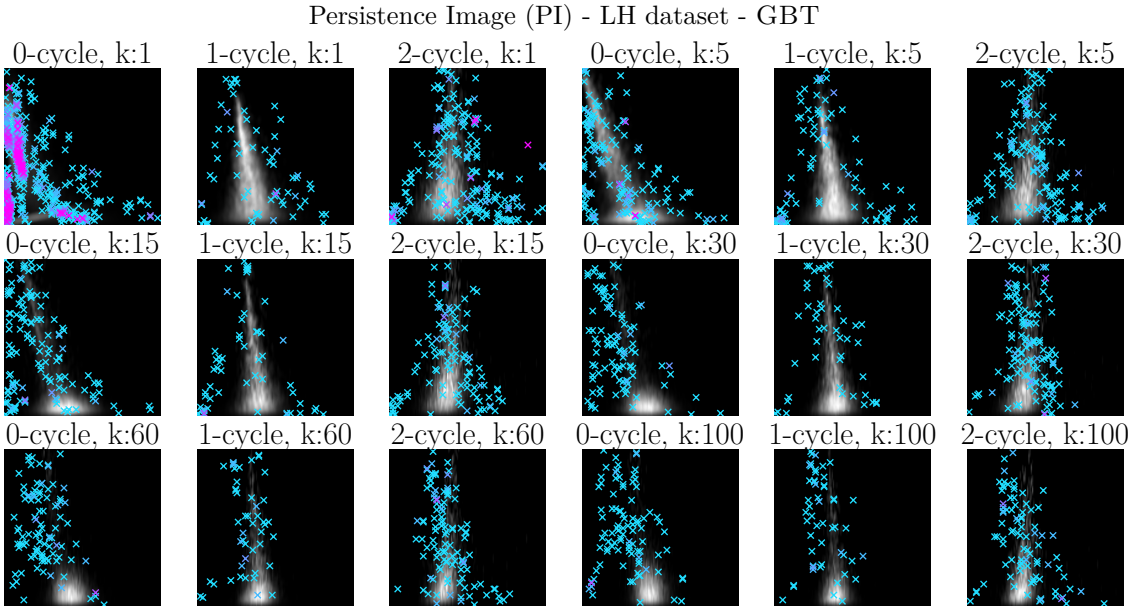

Figure 9. Feature importance map for the GBT model trained on the LH dataset, focused exclusively on $\Omega_{\rm m}$. Crosses indicate the positions of non-null features, with cyan marking pixels where the feature weight is non-null and below 1 and pink denoting weights above 1. The background shows the persistence image for a realization with $\Omega_{\rm m} = 0.3227$. Each image represents a different input channel, corresponding to all cycles with varying k-neighbors in the filtration.

JCAP09(2025)064

methods (neural-based and tree-based) to assess their strengths and limitations in extracting fundamental cosmological parameters and primordial non-Gaussianity amplitudes.

Our results demonstrate that persistence images are highly effective for recovering cosmological parameters under various scenarios. The information contained in persistence images consistently outperformed other methods for parameters such as $\Omega_{\rm m}$ and σ_8, achieving higher R^2 values and lower RMSE compared to the PS/BS data. Notably, persistence images were particularly effective in constraining $f_{\rm NL}^{\rm loc}$. Interestingly, GBT provided a computationally cheaper alternative and, for $f_{\rm NL}^{\rm loc}$, outperformed neural networks when trained on the same dataset.

As highlighted in [77], persistent homology reveals clustering differences that are not captured by second-order statistics or local density alone, but instead reflect the higher-order connectivity of the cosmic web. Because cosmological models imprint subtle changes not only in how much structure forms but also in how that structure connects and surrounds matter, topological measures can detect morphological differences that traditional summary statistics often compress, making them potentially more sensitive and complementary for cosmological parameter inference. This is particularly relevant for parameters such as $\Omega_{\rm m}$, σ_8, and $f_{\rm NL}^{\rm loc}$, which physically influence both the halo distribution and the surrounding matter field.

Combining PI and PS/BS data in the Hybrid model resulted in only marginal improvements over the CNN alone. This suggests that the PS/BS do not contain additional information relevant to parameter inference beyond what is already present in the PIs, nor do they offer complementary information. However, our analysis is limited by the properties of the available datasets: they are relatively small from a machine learning standpoint, which forces us to limit the flexibility of the models considered to prevent overfitting. The different LHs considered here also have different properties in terms of number of parameters varied. It would be interesting to explore the robustness of the PI summary statistic to marginalization over other cosmological parameters.

Despite these successes, all models struggled to constrain $f_{\rm NL}^{\rm equi}$. This may be because the LH_$f_{\rm NL}^{\rm equi}$ dataset varies several cosmological parameters over a large range. It would also be interesting to explore whether this statistic contains information about $f_{\rm NL}^{\rm equi}$ by including tighter priors on cosmological parameters such as $\Omega_{\rm m}$.

Parameters such as $f_{\rm NL}^{\rm loc}$, $\Omega_{\rm m}$, and σ_8 tend to increase the number of halos and amplify clustering features that are more readily captured by persistence images. In contrast, $f_{\rm NL}^{\rm equi}$ has a subtler effect, and improving its constraints would require larger simulation volumes to reduce sample variance. Notably, persistent homology offers a distinct advantage by focusing on small-scale features that remain sensitive to equilateral non-Gaussianity without relying on large-scale modes. Increasing simulation volumes could significantly enhance constraints on $f_{\rm NL}^{\rm equi}$ by enabling the detection of more features across multiple filtration scales. It was argued in [62] that this can be achieved by joining many small simulations since PIs don't necessarily derive their constraining power from the large scales.

While direct comparisons to other studies are complicated by differences in dataset preparation — such as variations in power spectrum/bispectrum measurements, redshift ranges, mass cuts, and Fourier binning — our findings remain broadly consistent with related work [26, 33, 52, 80, 97–103].

JCAP09(2025)064

As shown in [77], persistent homology is sensitive to the halo mass range, with more massive halos typically tracing larger-scale topological features than their lower-mass counterparts. Future work should explicitly evaluate how variations in halo populations influence the constraining power of topological summary statistics. Moreover, our feature importance analyses with `XGBoost` indicate that the model primarily focuses on early-born topological features, those that emerge early in the filtration process. In our setup, these early features correspond to haloes embedded in high-density regions, which are often associated with massive haloes. This suggests that the model may be capturing parameter dependencies through mass-related topological imprints. Given that cosmological parameters such as $\Omega_{\rm m}$ and $f_{\rm NL}^{\rm loc}$ strongly affect the abundance of high-mass haloes, we anticipate that different halo populations will vary in their constraining power, with some providing stronger sensitivity to specific cosmological parameters than others.

Direct inference using persistence diagrams through methods such as `DeepSets` or `PersLay` offers a promising direction for future research [104, 105]. While substantial performance gains are not guaranteed, these approaches could reveal complementary information and improve our understanding of how persistence features encode cosmological information.

Transitioning to simulation-based inference frameworks, such as those outlined in recent studies [98, 106–115], could further enhance parameter estimation by enabling direct sampling of posterior distributions. Successfully integrating these methods with convolutional neural networks or tree-based models like `XGBoost` will require careful development and experimentation, but they hold significant potential for advancing the field of cosmological inference.

Acknowledgments

We thank Moritz Münchmeyer for useful discussions and access to his group's computational resources in the early stages of this project. We thank Matteo Biagetti, Mathieu Carrière, and Francisco Villaescusa-Navarro for useful discussions and feedback on this project. The work of J.H.T.Y. and G.S. is supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics under Award Numbers DE-SC-0023719 and DE-SC-0017647. J.N. is supported by FONDECYT Regular grant 1211545. J.C. is supported by FONDECYT de Postdoctorado, N° 3240444. Powered@NLHPC: this research was partially supported by the supercomputing infrastructure of NLHPC (CCSS210001). G.C. acknowledges support from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355.

We ran test simulations, all persistent homology calculations, and power spectrum and bispectrum measurements using the computing resources and assistance of the University of Wisconsin-Madison Center for High Throughput Computing (CHTC) [116] in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the OSG Consortium, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

JCAP09(2025)064

References

[1] J.H.T. Yip, A. Rouhiainen and G. Shiu, *Learning from Topology: Cosmological Parameter Estimation from the Large-scale Structure*, *Mach. Learn. Sci. Tech.* **6** (2025) 025063 [arXiv:2308.02636] [INSPIRE].

[2] SPHEREx collaboration, *Cosmology with the SPHEREX All-Sky Spectral Survey*, arXiv:1412.4872 [INSPIRE].

[3] L. Amendola et al., *Cosmology and fundamental physics with the Euclid satellite*, *Living Rev. Rel.* **21** (2018) 2 [arXiv:1606.00180] [INSPIRE].

[4] H. Zhan and J.A. Tyson, *Cosmology with the Large Synoptic Survey Telescope: an Overview*, *Rept. Prog. Phys.* **81** (2018) 066901 [arXiv:1707.06948] [INSPIRE].

[5] M.M. Ivanov, M. Simonović and M. Zaldarriaga, *Cosmological Parameters from the BOSS Galaxy Power Spectrum*, *JCAP* **05** (2020) 042 [arXiv:1909.05277] [INSPIRE].

[6] P. Zhang et al., *BOSS Correlation Function analysis from the Effective Field Theory of Large-Scale Structure*, *JCAP* **02** (2022) 036 [arXiv:2110.07539] [INSPIRE].

[7] M.M. Ivanov, M. Simonović and M. Zaldarriaga, *Cosmological Parameters and Neutrino Masses from the Final Planck and Full-Shape BOSS Data*, *Phys. Rev. D* **101** (2020) 083504 [arXiv:1912.08208] [INSPIRE].

[8] DES collaboration, *Dark Energy Survey Year 1 Results: Cosmological constraints from cluster abundances and weak lensing*, *Phys. Rev. D* **102** (2020) 023509 [arXiv:2002.11124] [INSPIRE].

[9] DES collaboration, *Dark Energy Survey Year 1 Results: Cosmological Constraints from Cluster Abundances, Weak Lensing, and Galaxy Correlations*, *Phys. Rev. Lett.* **126** (2021) 141301 [arXiv:2010.01138] [INSPIRE].

[10] A. Vikhlinin et al., *Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints*, *Astrophys. J.* **692** (2009) 1060 [arXiv:0812.2720] [INSPIRE].

[11] DSDD collaboration, *Cosmological Constraints from the SDSS maxBCG Cluster Catalog*, *Astrophys. J.* **708** (2010) 645 [arXiv:0902.3702] [INSPIRE].

[12] SPT and DES collaborations, *SPT clusters with DES and HST weak lensing. II. Cosmological constraints from the abundance of massive halos*, *Phys. Rev. D* **110** (2024) 083510 [arXiv:2401.02075] [INSPIRE].

[13] V. Ghirardini et al., *The SRG/eROSITA all-sky survey — Cosmology constraints from cluster abundances in the western Galactic hemisphere*, *Astron. Astrophys.* **689** (2024) A298 [arXiv:2402.08458] [INSPIRE].

[14] eBOSS collaboration, *Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory*, *Phys. Rev. D* **103** (2021) 083533 [arXiv:2007.08991] [INSPIRE].

[15] L. Verde et al., *The 2dF Galaxy Redshift Survey: The Bias of galaxies and the density of the Universe*, *Mon. Not. Roy. Astron. Soc.* **335** (2002) 432 [astro-ph/0112161] [INSPIRE].

[16] H. Gil-Marín et al., *The power spectrum and bispectrum of SDSS DR11 BOSS galaxies — I. Bias and gravity*, *Mon. Not. Roy. Astron. Soc.* **451** (2015) 539 [arXiv:1407.5668] [INSPIRE].

[17] G. D'Amico et al., *The Cosmological Analysis of the SDSS/BOSS data from the Effective Field Theory of Large-Scale Structure*, *JCAP* **05** (2020) 005 [arXiv:1909.05271] [INSPIRE].

JCAP09(2025)064

[18] G. D'Amico et al., *The BOSS bispectrum analysis at one loop from the Effective Field Theory of Large-Scale Structure*, *JCAP* **05** (2024) 059 [arXiv:2206.08327] [INSPIRE].

[19] M.M. Ivanov et al., *Cosmology with the galaxy bispectrum multipoles: Optimal estimation and application to BOSS data*, *Phys. Rev. D* **107** (2023) 083515 [arXiv:2302.04414] [INSPIRE].

[20] T. Colas et al., *Efficient Cosmological Analysis of the SDSS/BOSS data from the Effective Field Theory of Large-Scale Structure*, *JCAP* **06** (2020) 001 [arXiv:1909.07951] [INSPIRE].

[21] M. Biagetti, *The Hunt for Primordial Interactions in the Large Scale Structures of the Universe*, *Galaxies* **7** (2019) 71 [arXiv:1906.12244] [INSPIRE].

[22] G. Cabass et al., *Constraints on multifield inflation from the BOSS galaxy survey*, *Phys. Rev. D* **106** (2022) 043506 [arXiv:2204.01781] [INSPIRE].

[23] G. Cabass et al., *Constraints on Single-Field Inflation from the BOSS Galaxy Survey*, *Phys. Rev. Lett.* **129** (2022) 021301 [arXiv:2201.07238] [INSPIRE].

[24] G. D'Amico, M. Lewandowski, L. Senatore and P. Zhang, *Limits on primordial non-Gaussianities from BOSS galaxy-clustering data*, *Phys. Rev. D* **111** (2025) 063514 [arXiv:2201.11518] [INSPIRE].

[25] M. White, *A marked correlation function for constraining modified gravity models*, *JCAP* **11** (2016) 057 [arXiv:1609.08632] [INSPIRE].

[26] G. Jung et al., *Quijote-PNG: Optimizing the Summary Statistics to Measure Primordial Non-Gaussianity*, *Astrophys. J.* **976** (2024) 109 [arXiv:2403.00490] [INSPIRE].

[27] E. Massara et al., *Using the Marked Power Spectrum to Detect the Signature of Neutrinos in Large-Scale Structure*, *Phys. Rev. Lett.* **126** (2021) 011301 [arXiv:2001.11024] [INSPIRE].

[28] M. Marinucci et al., *The constraining power of the Marked Power Spectrum: an analytical study*, arXiv:2411.14377 [INSPIRE].

[29] E. Massara et al., *Cosmological Information in the Marked Power Spectrum of the Galaxy Field*, *Astrophys. J.* **951** (2023) 70 [arXiv:2206.01709] [INSPIRE].

[30] J.A. Cowell, D. Alonso and J. Liu, *Optimizing marked power spectra for cosmology*, *Mon. Not. Roy. Astron. Soc.* **535** (2024) 3129 [arXiv:2409.05695] [INSPIRE].

[31] J. Hou et al., *Cosmological constraints from the redshift-space galaxy skew spectra*, *Phys. Rev. D* **109** (2024) 103528 [arXiv:2401.15074] [INSPIRE].

[32] M. Schmittfull and A. Moradinezhad Dizgah, *Galaxy skew-spectra in redshift-space*, *JCAP* **03** (2021) 020 [arXiv:2010.14267] [INSPIRE].

[33] M. Peron, G. Jung, M. Liguori and M. Pietroni, *Constraining primordial non-Gaussianity from large scale structure with the wavelet scattering transform*, *JCAP* **07** (2024) 021 [arXiv:2403.17657] [INSPIRE].

[34] M. Eickenberg et al., *Wavelet Moments for Cosmological Parameter Estimation*, arXiv:2204.07646 [INSPIRE].

[35] G. Valogiannis and C. Dvorkin, *Going beyond the galaxy power spectrum: An analysis of BOSS data with wavelet scattering transforms*, *Phys. Rev. D* **106** (2022) 103509 [arXiv:2204.13717] [INSPIRE].

[36] G. Valogiannis, S. Yuan and C. Dvorkin, *Precise cosmological constraints from BOSS galaxy clustering with a simulation-based emulator of the wavelet scattering transform*, *Phys. Rev. D* **109** (2024) 103503 [arXiv:2310.16116] [INSPIRE].

JCAP09(2025)064

[37] C. Uhlemann et al., *Fisher for complements: Extracting cosmology and neutrino mass from the counts-in-cells PDF*, *Mon. Not. Roy. Astron. Soc.* **495** (2020) 4006 [arXiv:1911.11158] [INSPIRE].

[38] O. Friedrich et al., *Primordial non-Gaussianity without tails — how to measure fNL with the bulk of the density PDF*, *Mon. Not. Roy. Astron. Soc.* **498** (2020) 464 [arXiv:1912.06621] [INSPIRE].

[39] B.M.C. Gould, L. Castiblanco, C. Uhlemann and O. Friedrich, *Cosmology on point: modelling spectroscopic tracer one-point statistics*, arXiv:2409.18182 [DOI:10.33232/001c.127800] [INSPIRE].

[40] G. D'Amico, M. Musso, J. Noreña and A. Paranjape, *Excursion Sets and Non-Gaussian Void Statistics*, *Phys. Rev. D* **83** (2011) 023521 [arXiv:1011.1229] [INSPIRE].

[41] M. Kamionkowski, L. Verde and R. Jimenez, *The Void Abundance with Non-Gaussian Primordial Perturbations*, *JCAP* **01** (2009) 010 [arXiv:0809.0506] [INSPIRE].

[42] A. Pisani et al., *Cosmic voids: a novel probe to shed light on our Universe*, arXiv:1903.05161 [INSPIRE].

[43] A. Banerjee and T. Abel, *Nearest neighbour distributions: New statistical measures for cosmological clustering*, *Mon. Not. Roy. Astron. Soc.* **500** (2020) 5479 [arXiv:2007.13342] [INSPIRE].

[44] W.R. Coulton, T. Abel and A. Banerjee, *Small-scale signatures of primordial non-Gaussianity in k-nearest neighbour cumulative distribution functions*, *Mon. Not. Roy. Astron. Soc.* **534** (2024) 1621 [arXiv:2309.15151] [INSPIRE].

[45] M. Lippich and A.G. Sánchez, *medusa: Minkowski functionals estimated from Delaunay tessellations of the three-dimensional large-scale structure*, *Mon. Not. Roy. Astron. Soc.* **508** (2021) 3771 [arXiv:2012.08529] [INSPIRE].

[46] W. Liu, A. Jiang and W. Fang, *Probing massive neutrinos with the Minkowski functionals of the galaxy distribution*, *JCAP* **09** (2023) 037 [arXiv:2302.08162] [INSPIRE].

[47] A. Jiang et al., *Minkowski functionals of large-scale structure as a probe of modified gravity*, *Phys. Rev. D* **109** (2024) 083537 [arXiv:2305.04520] [INSPIRE].

[48] SIMBIG collaboration, *Field-level simulation-based inference of galaxy clustering with convolutional neural networks*, *Phys. Rev. D* **109** (2024) 083536 [arXiv:2310.15256] [INSPIRE].

[49] H. Shao et al., *Robust Field-level Inference of Cosmological Parameters with Dark Matter Halos*, *Astrophys. J.* **944** (2023) 27 [arXiv:2209.06843] [INSPIRE].

[50] N.S.M. de Santi et al., *Robust Field-level Likelihood-free Inference with Galaxies*, *Astrophys. J.* **952** (2023) 69 [arXiv:2302.14101] [INSPIRE].

[51] S. Anagnostidis et al., *Cosmology from Galaxy Redshift Surveys with PointNet*, arXiv:2211.12346 [INSPIRE].

[52] A. Chatterjee and F. Villaescusa-Navarro, *Cosmology from Point Clouds with Dark Matter Halos from the Quijote Simulations*, *Astrophys. J.* **985** (2025) 132 [arXiv:2405.13119] [INSPIRE].

[53] A. Barreira, T. Lazeyras and F. Schmidt, *Galaxy bias from forward models: linear and second-order bias of IllustrisTNG galaxies*, *JCAP* **08** (2021) 029 [arXiv:2105.02876] [INSPIRE].

[54] N.-M. Nguyen et al., *How Much Information Can Be Extracted from Galaxy Clustering at the Field Level?*, *Phys. Rev. Lett.* **133** (2024) 221006 [arXiv:2403.03220] [INSPIRE].

JCAP09(2025)064

[55] I. Babić, F. Schmidt and B. Tucci, *Straightening the Ruler: Field-Level Inference of the BAO Scale with LEFTfield*, arXiv:2407.01524 [INSPIRE].

[56] A. Kostić, N.-M. Nguyen, F. Schmidt and M. Reinecke, *Consistency tests of field level inference with the EFT likelihood*, *JCAP* **07** (2023) 063 [arXiv:2212.07875] [INSPIRE].

[57] G. Wilding et al., *Persistent homology of the cosmic web — I. Hierarchical topology in ΛCDM cosmologies*, *Mon. Not. Roy. Astron. Soc.* **507** (2021) 2968 [arXiv:2011.12851] [INSPIRE].

[58] M.A. Aragón-Calvo, R. van de Weygaert and B.J.T. Jones, *Multiscale Phenomenology of the Cosmic Web*, *Mon. Not. Roy. Astron. Soc.* **408** (2010) 2163 [arXiv:1007.0742] [INSPIRE].

[59] T. Sousbie, C. Pichon and H. Kawahara, *The persistent cosmic web and its filamentary structure II: Illustrations*, *Mon. Not. Roy. Astron. Soc.* **414** (2011) 384 [arXiv:1009.4014] [INSPIRE].

[60] P. Pranav et al., *The Topology of the Cosmic Web in Terms of Persistent Betti Numbers*, *Mon. Not. Roy. Astron. Soc.* **465** (2017) 4281 [arXiv:1608.04519] [INSPIRE].

[61] A. Cole, M. Biagetti and G. Shiu, *Topological Echoes of Primordial Physics in the Universe at Large Scales*, in the proceedings of the *34th Conference on Neural Information Processing Systems*, Online Conference, Canada, 06–12 December 2020 [arXiv:2012.03616] [INSPIRE].

[62] M. Biagetti et al., *Fisher forecasts for primordial non-Gaussianity from persistent homology*, *JCAP* **10** (2022) 002 [arXiv:2203.08262] [INSPIRE].

[63] J.H.T. Yip et al., *Cosmology with persistent homology: a Fisher forecast*, *JCAP* **09** (2024) 034 [arXiv:2403.13985] [INSPIRE].

[64] M. Biagetti, A. Cole and G. Shiu, *The Persistence of Large Scale Structures I: Primordial non-Gaussianity*, *JCAP* **04** (2021) 061 [arXiv:2009.04819] [INSPIRE].

[65] J. Feldbrugge et al., *Stochastic Homology of Gaussian vs. non-Gaussian Random Fields: Graphs towards Betti Numbers and Persistence Diagrams*, *JCAP* **09** (2019) 052 [arXiv:1908.01619] [INSPIRE].

[66] S. Heydenreich, B. Brück and J. Harnois-Déraps, *Persistent homology in cosmic shear: constraining parameters with topological data analysis*, *Astron. Astrophys.* **648** (2021) A74 [arXiv:2007.13724] [INSPIRE].

[67] S. Heydenreich et al., *Persistent homology in cosmic shear — II. A tomographic analysis of DES-Y1*, *Astron. Astrophys.* **667** (2022) A125 [arXiv:2204.11831] [INSPIRE].

[68] M.H.J. Kanafi, S. Ansarifard and S.M.S. Movahed, *Imprint of massive neutrinos on Persistent Homology of large-scale structure*, arXiv:2311.13520 [INSPIRE].

[69] H. Edelsbrunner and J. Harer, *Computational Topology — an Introduction*, American Mathematical Society (2010).

[70] G. Carlsson and M. Vejdemo-Johansson, *Topological Data Analysis with Applications*, Topological Data Analysis with Applications, Cambridge University Press (2021) [DOI:10.1017/9781108975704].

[71] L. Wasserman, *Topological Data Analysis*, arXiv:1609.08227.

[72] H. Edelsbrunner and E.P. Mücke, *Three-dimensional alpha shapes*, *ACM Trans. Graph.* **13** (1994) 43.

[73] H. Edelsbrunner, D. Kirkpatrick and R. Seidel, *On the shape of a set of points in the plane*, *IEEE Trans. Inform. Theory* **29** (1983) 551.

JCAP09(2025)064

[74] The GUDHI Project, *GUDHI User and Reference Manual*, GUDHI Editorial Board, 3.11.0 ed. (2025), https://gudhi.inria.fr/doc/3.11.0/.

[75] The CGAL Project, *CGAL User and Reference Manual*, CGAL Editorial Board, 4.14 ed. (2019), https://doc.cgal.org/4.14/Manual/packages.html.

[76] R. van de Weygaert et al., *Alpha, Betti and the Megaparsec Universe: on the Topology of the Cosmic Web*, *Trans. Comput. Sci.* **14** (2011) 60 [arXiv:1306.3640] [INSPIRE].

[77] R. Bermejo et al., *Topological bias: how haloes trace structural patterns in the cosmic web*, *Mon. Not. Roy. Astron. Soc.* **529** (2024) 4325 [arXiv:2206.14655] [INSPIRE].

[78] F. Villaescusa-Navarro et al., *The Quijote simulations*, *Astrophys. J. Suppl.* **250** (2020) 2 [arXiv:1909.05273] [INSPIRE].

[79] W.R. Coulton et al., *Quijote-PNG: Simulations of Primordial Non-Gaussianity and the Information Content of the Matter Field Power Spectrum and Bispectrum*, *Astrophys. J.* **943** (2023) 64 [arXiv:2206.01619] [INSPIRE].

[80] G. Jung et al., *Quijote-PNG: The Information Content of the Halo Mass Function*, *Astrophys. J.* **957** (2023) 50 [arXiv:2305.10597] [INSPIRE].

[81] E. Sefusatti, M. Crocce, R. Scoccimarro and H. Couchman, *Accurate Estimators of Correlation Functions in Fourier Space*, *Mon. Not. Roy. Astron. Soc.* **460** (2016) 3624 [arXiv:1512.07295] [INSPIRE].

[82] N. Jeffrey and B.D. Wandelt, *Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks*, in the proceedings of the *34th Conference on Neural Information Processing Systems*, Online Conference, Canada, 06–12 December 2020 [arXiv:2011.05991] [INSPIRE].

[83] F. Villaescusa-Navarro et al., *Multifield Cosmology with Artificial Intelligence*, arXiv:2109.09747 [INSPIRE].

[84] F. Villaescusa-Navarro et al., *Robust marginalization of baryonic effects for cosmological inference at the field level*, arXiv:2109.10360 [INSPIRE].

[85] P. Villanueva-Domingo et al., *Inferring Halo Masses with Graph Neural Networks*, *Astrophys. J.* **935** (2022) 30 [arXiv:2111.08683] [INSPIRE].

[86] B.Y. Wang, A. Pisani, F. Villaescusa-Navarro and B.D. Wandelt, *Machine-learning Cosmology from Void Properties*, *Astrophys. J.* **955** (2023) 131 [arXiv:2212.06860] [INSPIRE].

[87] P. Villanueva-Domingo and F. Villaescusa-Navarro, *Learning Cosmology and Clustering with Cosmic Graphs*, *Astrophys. J.* **937** (2022) 115 [arXiv:2204.13713] [INSPIRE].

[88] L.A. Perez et al., *Constraining Cosmology with Machine Learning and Galaxy Clustering: The CAMELS-SAM Suite*, *Astrophys. J.* **954** (2023) 11 [arXiv:2204.02408] [INSPIRE].

[89] F. Villaescusa-Navarro et al., *Cosmology with One Galaxy?*, *Astrophys. J.* **929** (2022) 132 [arXiv:2201.02202] [INSPIRE].

[90] C. Chawak et al., *Cosmology with Multiple Galaxies*, *Astrophys. J.* **969** (2024) 105 [arXiv:2309.12048] [INSPIRE].

[91] N.S.M. de Santi et al., *Field-level simulation-based inference with galaxy catalogs: the impact of systematic effects*, *JCAP* **01** (2025) 082 [arXiv:2310.15234] [INSPIRE].

[92] Y. Gondhalekar and K. Moriwaki, *Convolutional Vision Transformer for Cosmology Parameter Inference*, arXiv:2411.14392 [INSPIRE].

JCAP09(2025)064

[93] L. Grinsztajn, E. Oyallon and G. Varoquaux, *Why do tree-based models still outperform deep learning on tabular data?*, *Adv. Neural Inf. Process. Syst.* **35** (2022) 507 [arXiv:2207.08815].

[94] A. Lazanu, *Extracting cosmological parameters from N-body simulations using machine learning techniques*, *JCAP* **09** (2021) 039 [arXiv:2106.11061] [INSPIRE].

[95] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, arXiv:1603.02754 [DOI:10.1145/2939672.2939785] [INSPIRE].

[96] D. Babich, P. Creminelli and M. Zaldarriaga, *The Shape of non-Gaussianities*, *JCAP* **08** (2004) 009 [astro-ph/0405356] [INSPIRE].

[97] G. Valogiannis and C. Dvorkin, *Towards an optimal estimation of cosmological parameters with the wavelet scattering transform*, *Phys. Rev. D* **105** (2022) 103534 [arXiv:2108.07821] [INSPIRE].

[98] E. Massara et al., SIMBIG*: Cosmological Constraints using Simulation-Based Inference of Galaxy Clustering with Marked Power Spectra*, arXiv:2404.04228 [INSPIRE].

[99] T.L. Makinen et al., *The Cosmic Graph: Optimal Information Extraction from Large-Scale Structure using Catalogues*, arXiv:2207.05202 [DOI:10.21105/astro.2207.05202] [INSPIRE].

[100] H.J. Hortúa, L.Á. García and L. Castañeda C., *Constraining cosmological parameters from N-body simulations with variational Bayesian neural networks*, *Front. Astron. Space Sci.* **10** (2023) 1139120 [arXiv:2301.03991] [INSPIRE].

[101] M. Ho et al., *LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology*, *Open J. Astrophys.* **7** (2024) 001c.120559 [arXiv:2402.05137] [INSPIRE].

[102] C. Cuesta-Lazaro and S. Mishra-Sharma, *Point cloud approach to generative modeling for galaxy surveys at the field level*, *Phys. Rev. D* **109** (2024) 123531 [arXiv:2311.17141] [INSPIRE].

[103] Z. Min et al., *Deep learning for cosmological parameter inference from a dark matter halo density field*, *Phys. Rev. D* **110** (2024) 063531 [arXiv:2404.09483] [INSPIRE].

[104] M. Zaheer et al., *Deep Sets*, *Adv. Neural Inf. Process. Syst.* **30** (2017) [arXiv:1703.06114] [INSPIRE].

[105] M. Carrière et al., *PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures*, arXiv:1904.09378.

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

[107] M. Reza et al., *Constraining Cosmology with Simulation-based inference and Optical Galaxy Cluster Abundance*, arXiv:2409.20507 [INSPIRE].

[108] B.K. Miller et al., *Truncated Marginal Neural Ratio Estimation*, in the proceedings of the *35th Conference on Neural Information Processing Systems*, Online Conference, Canada, 06–14 December 2021 [DOI:10.5281/zenodo.5043706] [arXiv:2107.01214] [INSPIRE].

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

[110] T.L. Makinen et al., *Hybrid Summary Statistics*, arXiv:2410.07548 [INSPIRE].

[111] G. Khullar et al., *DIGS: deep inference of galaxy spectra with neural posterior estimation*, *Mach. Learn. Sci. Tech.* **3** (2022) 04LT04 [arXiv:2211.09126] [INSPIRE].

JCAP09(2025)064

[112] K. Lehman, S. Krippendorf, J. Weller and K. Dolag, *Learning Optimal and Interpretable Summary Statistics of Galaxy Catalogs with SBI*, arXiv:2411.08957 [INSPIRE].

[113] A. Saxena et al., *Simulation-based inference of the sky-averaged 21-cm signal from CD-EoR with REACH*, *RAS Tech. Instrum.* **3** (2024) 724 [arXiv:2403.14618] [INSPIRE].

[114] C.H. Hahn et al., *A forward modeling approach to analyzing galaxy clustering with* SIMBIG, *Proc. Nat. Acad. Sci.* **120** (2023) e2218810120 [arXiv:2211.00723] [INSPIRE].

[115] B. Tucci and F. Schmidt, *EFTofLSS meets simulation-based inference: σ_8 from biased tracers*, *JCAP* **05** (2024) 063 [arXiv:2310.03741] [INSPIRE].

[116] Center for High Throughput Computing, *Center for high throughput computing*, DOI:10.21231/GNT1-HW21 (2006).

JCAP09(2025)064