Differentiating small-scale subhalo distributions in CDM and WDM models using persistent homology

Jessi Cisewski-Kehe*
Department of Statistics, University of Wisconsin-Madison, 1300 University Ave, Madison, WI 53706, USA

Brittany Terese Fasy
Gianforte School of Computing, Montana State University, 357 Barnard Hall, P.O. Box 173880, Bozeman, MT, 59717-3880, USA

Wojciech Hellwing and Paweł Drozda
Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland

Mark R. Lovell
Science Institute, University of Iceland, Dunhaga 5, 107 Reykjavík

Mike Wu
Computer Science Department, Stanford University, 353 Jane Stanford Way, Stanford, CA, 94305, USA
(Dated: August 25, 2022)

The spatial distribution of galaxies at sufficiently small scales will encode information about the identity of the dark matter. We develop a novel description of the halo distribution using persistent homology summaries, in which collections of points are decomposed into clusters, loops and voids. We apply these methods, together with a set of hypothesis tests, to dark matter haloes in MW-analog environment regions of the cold dark matter (CDM) and warm dark matter (WDM) Copernicus Complexio N-body cosmological simulations. The results of the hypothesis tests find statistically significant differences (p-values ≤ 0.001) between the CDM and WDM structures, and the functional summaries of persistence diagrams detect differences at scales that are distinct from the comparison spatial point process functional summaries considered (including the two-point correlation function). The differences between the models are driven most strongly at filtration scales ~ 100 kpc, where CDM generates larger numbers of unconnected halo clusters while WDM instead generates loops. This study was conducted on dark matter haloes generally; future work will involve applying the same methods to realistic galaxy catalogues.

I. INTRODUCTION

The large scale structure (LSS)—as defined by the spatial distribution of galaxies—encodes information on many vital aspects of the standard model of cosmology that remain open questions in physics [1–4]. For example, the LSS is sensitive to the characteristics of dark energy, the unexplained phenomenon that drives the accelerated expansion of the Universe [5, 6] and also holds clues as to the nature of dark matter (DM). Typical LSS observables that are relevant for DM studies include the abundance of low mass galaxies [7, 8], the paucity of galaxies in voids [9] and the spatial distribution of MW satellite galaxies [10]. An additional, as yet largely untapped, method for analysing LSS models is the application of topological methods to the distribution of galaxies and haloes. These methods describe the spatial distribution of points as different dimensional holes with clusters, filaments loops, and voids in dimensions 0, 1, and 2, respectively, and it is possible to envisage that the imprint of DM physics on the primordial density field may be detectable in their topological statistics [5, 11]. In this paper we will apply topological methods in order to identify differences between two competing DM models. The simplest viable model of DM is the cold DM matter model (CDM), in which the DM particle has a negligible velocity dispersion at early times and thus DM halos are able to start collapsing early and in large quantities. The combination of CDM with the cosmological constant model of dark energy is known as ΛCDM. This model has enjoyed success in predicting the properties of the cosmic microwave background (CMB) radiation [12] and the distribution of galaxies at large scales (>2 Mpc) [13]. However, at smaller scales (<1 Mpc) there are tensions among others with the densities of dwarf galaxies that may hint at problems for the CDM model [3]. Given the simultaneous failure to detect the particle physics candidates that correspond to CDM in direct detection experiments [14, 15] or in indirect detection observations [16], it is important to consider alternatives.

One compelling alternative to CDM is the warm DM (WDM) model, in which the DM particles have a significant velocity dispersion in the early Universe [17]. The

* jjkehe@wisc.edu

effects of this velocity dispersion include a drastic reduction in the number of low mass DM halos. In this study we compare simulations of these two models to determine whether persistent homology can detect differences in the DM halo spatial distribution. In Fig. 1, we present images of two realizations of the Copernicus Complexio (COCO) cosmological volume [18], one simulated with the CDM model, and the second with the WDM model [19]. The large scale distribution of matter is nearly identical in the two images—thus WDM preserves the large scale successes in explaining the distribution of massive galaxies of CDM—but at smaller scales the abundance of WDM halos is strongly suppressed relative to CDM, and the distribution of the remaining subhalos is much less homogeneous.

In this work, we investigate differences in the spatial distribution of DM haloes as described in CDM and WDM. The primary goal is to ascertain whether topological methods are sensitive to differences between the models and the second goal is to interpret the differences to determine whether topological methods have the potential to discern which DM model most accurately describes the properties of our own Universe. We restrict our analysis to the distribution of haloes, which will work as a proof of concept. A comprehensive comparison with observations will require a mock galaxy catalogue and we defer this step to future work.

The persistent homology formulation of topology offers a novel way to represent, visualize, and interpret complex data by extracting homological features, which can be used to infer properties of the underlying structures. Homological features include the decomposition of halo distributions into clusters, filaments loops, and voids at different scales controlled by a parameter that is analogous to halo linking lengths—which in statistics is known as a filtration parameter—and persistent homology in particular tracks how the number of such features changes as the filtration parameter is increased. It has been successfully applied to problems in astronomy (e.g., Refs. [5, 20–27]), along with other areas of science (e.g., Refs. Ref. [28–31]). There have been proposals for hypothesis testing using persistent homology (e.g., Refs. [31–35]), which we build on as we construct tests that can detect differences between DM model predictions in the LSS.

We investigate several test statistics to discriminate between the CDM and WDM halo spatial distributions that are based on persistent homology functional summaries. Each functional summary is a different transformation of information to a function that approximates a property of the topological features, and is a function of the filtration parameter. We also consider different visualizations in order to investigate detected differences.

This paper is organised as follows. We begin with background on the cosmological simulation data we use in the analysis (§II), then we introduce persistent homology and functional summaries of persistence diagrams that are used in the proposed test statistics (§III). Then the hypothesis testing framework is presented (§IV), followed by the investigation of the cosmological simulation data (§V). We end with concluding remarks (§VI).

II. COSMOLOGICAL SIMULATION DATA

This section begins with a description of the COCO simulations, and then continues with our procedure for selecting MW halo-analog sample regions.

A. The Copernicus Complexio (COCO) cosmological simulations

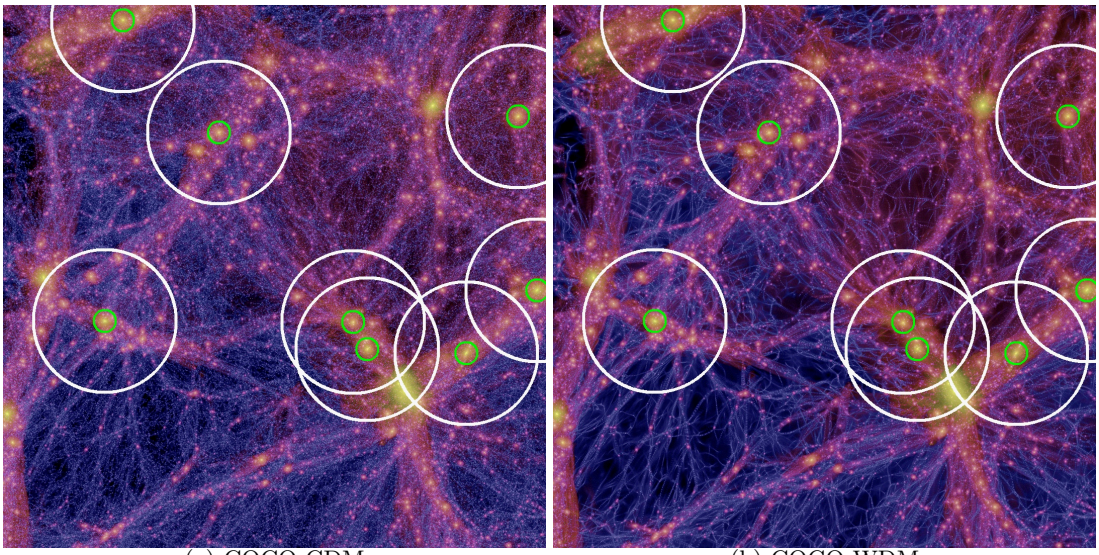
The COCO simulation volume constitutes a high resolution spherical region of space with a comoving radius of approximately 25 Mpc; the full (low-resolution) simulation volume is a periodic box 100 Mpc on a side.[1] The numerical integration of the gravitational forces begins at redshift 127. The cosmological parameters are consistent with the 7-year results from the WMAP satellites: matter density $\Omega_0 = 0.272$, dark energy density $\Omega_\Lambda = 0.728$, $\Omega_b = 0.04455$, Hubble parameter $h_0 = 0.704$, spectral index $n_s = 0.967$, and power spectrum normalization $\sigma_8 = 0.81$. The mass of the simulation particle is $1.135 \times 10^5\ M_\odot$. DM halos and subhalos were identified using the SUBFIND algorithm [36], and the smallest permitted number of particles to identfy a subhalo is 20 particles. Our definition of halo mass is the total mass bound gravitationally to each halo as determined by the halo finder.

Two copies of this volume were run, the first applying CDM [18] and the second WDM [19]. Both simulations use the same initial phases, and differ in that the WDM simulation had wave amplitudes rescaled using the transfer function of a 3.3 keV thermal relic DM particle, with the relic mass chosen to be in agreement with the Lyman-α forest constraints of Ref. [37]. This results in the suppression of structure on the scale of dwarf galaxies. These large-scale structure similarities between the WDM and CDM data due to the same initial phases are shown in Fig. 1. One issue peculiar to WDM simulations is the spurious numerical fragmentation of filaments into halos; these so-called spurious subhalos are identified and removed from the halo catalog using the algorithm described in Ref. [38].

B. Milky Way-analog DM halos and their associated halo samples

Given that we intend to use future work to compare the models with observations of galaxies around our own

[1] All distances are in comoving Mpc.



(a) COCO-CDM (b) COCO-WDM

FIG. 1: Illustration of the DM distributions in the COCO-CDM (a) and COCO-WDM (b) simulations. Each image is a slice through the simulation of 23 Mpc on a side with an image depth of 10 Mpc. The image intensity encodes the DM column density and the image color indicates velocity dispersion. Eight of the 77 volumes used in this study (see §II B) are included in this slice, and their locations are indicated as follows. The MW-analog halo on which each volume is centered is enclosed by a green circle, and the full extent of the analysis volume (a radius 3 Mpc) is shown with a white circle. Note that the apparent overlap of the white circles in this projection does not imply that the volumes overlap: there can still be considerable separation between the volumes in the depth direction. See §II for details on the COCO data.

MW, we identify MW-analog halos and their surrounding regions in the two simulations. The criteria for our MW-analog halos were that they must be located within 21 Mpc of the center of the simulation[2] and have a mass in the range $[0.5, 2] \times 10^{12}\ M_\odot$. We also required that there be no other halo with a mass greater than $0.5 \times 10^{12}\ M_\odot$ within 0.7 Mpc. This procedure resulted in 77 MW-analog DM halos in each of the COCO CDM and WDM simulations; for each MW-analog DM halo in the WDM data, there is a matching MW-analog DM halo in the CDM.

We now discuss our selection of halos in the vicinity of the 77 MW halo-analogs. In both CDM and WDM realizations we identify halos that are within 3 Mpc of the MW-analog[3]. For the WDM case we include all halos in the 3 Mpc region. However, CDM forms hundreds of times more halos than WDM in our resolved mass range. If we were to include all CDM halos, the abundance difference would dominate our statistical results. Therefore, the CDM samples were downsampled to match the number of DM halos in the corresponding WDM sample. The downsampling was accomplished by selecting the most massive DM halos from each of the CDM samples. An example of one of the MW-analog DM halo neighborhoods from COCO-CDM and COCO-WDM is displayed in Fig. 2.

The two sets of samples for the CDM and WDM data are defined as

$$\mathbb{Y}_c = \{\mathbf{Y}_{c,1}, \ldots, \mathbf{Y}_{c,77}\}, \mathbb{Y}_w = \{\mathbf{Y}_{w,1}, \ldots, \mathbf{Y}_{w,77}\} \quad (1)$$

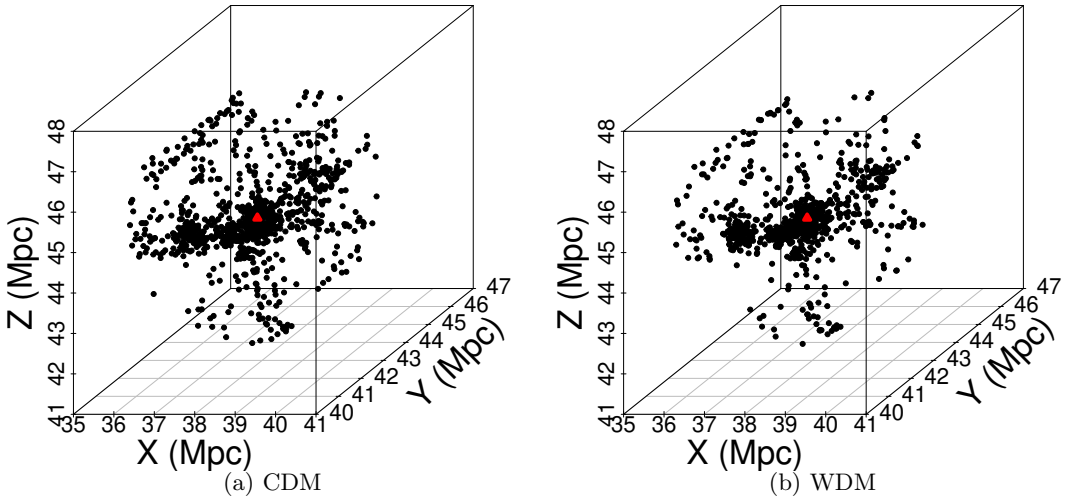
where $\mathbb{Y}_c$ and $\mathbb{Y}_w$ represent the set of 77 CDM and 77 WDM samples, respectively. Each $\mathbf{Y}_{k,i} \in \mathbb{R}^{n_i \times 3}$ for $k = c, w$ and $i = 1, \ldots, 77$ where n_i indicates the number of DM halos in sample i; an individual DM halo in simulation k of sample i is indicated by $Y_{k,i,j}$ for $j = 1, \ldots, n_i$.

III. TOPOLOGICAL DATA ANALYSIS METHODS FOR QUANTIFYING LSS

Homology is one way to study the features of topological spaces (e.g., manifolds); specifically, the multi-dimensional "holes" in the space (e.g., connected components, loops, voids). Persistent homology studies the spatial structure of a parameterized family of topological spaces that keeps track of the so-called *births* and *deaths* of homological features as a topological space changes with a filtration parameter. In particular, we focus on point cloud data, where each point can represent some unit of mass or an object (e.g., a point may represent

[2] The central high-resolution sphere of COCO extends out to about 25 Mpc.

[3] Objects with ∼0.2 Mpc of each analog are typically referred to as 'subhalos' that orbit within the analog 'host halo.' In this study we refer to all bound DM objects simply as 'halos' and include all of them in our analysis, not drawing any distinction between 'subhalos' and other classes of object.



(a) CDM (b) WDM

FIG. 2: A MW-analog DM halo neighborhood sample for (a) the COCO-CDM data and (b) the COCO-WDM data. The red triangles indicate the MW-analog DM halo that was selected along with the other DM halos that are within 3 Mpc from it.

a center of a DM halo). In this section, we provide a brief overview of the necessary concepts; however, see, e.g., Refs. [39–41] for a more thorough introduction to algebraic and computational topology. The homological features that are tracked in the filtration have cosmological interpretations in dimensions zero, one, and two. Before providing more details about persistent homology, we explain the interpretation of different dimensional holes with respect to the distribution of DM halos.

a. Clusters A *connected component*, or zeroth-dimensional homology feature (H_0), is a maximal subspace of a topological space that cannot be covered by two disjoint open sets; that is, a connected component is a *whole piece* of the space. For example, under some assumptions on the topological space, a connected component is a cluster of data points.[4] In cosmology, the connected components represent clusters of halos or galaxies. Persistent homology tracks the appearance of new connected components and the merging of distinct components.

b. Filaments and Loops A *loop*, or one-dimensional homology feature (H_1), provides information about the connectivity of data. As the filtration parameter increases, nearby connected components can merge together in such a way that a loop **or cycle** is formed. For DM halos, this would appear as filaments of halos joined together in a loop.

c. Cosmological Voids A *void*, or two-dimensional homology feature (H_2), represents the boundary of three-dimensional empty regions within the topological space (e.g., the boundary of a football). In cosmology, these are the thin walls surrounding the low-density regions that are typically referred to as cosmological voids.

[4] There is no clear or established relationship between the definition of homological clusters and galaxy clusters. In this paper, the term 'cluster' is only used for the homology definition.

A. Persistent homology

Persistent homology is a framework for ~~computing~~ **describing** the homology of a data set at different scales. Given a data set, one defines a filtration (that is, a sequence of nested topological spaces) of intermediate structures, on which the homology is computed at different values of the filtration parameter. ~~Homology group generators~~ **Homological features (specifically, the homology generators)** are tracked as they form and ~~die~~ **merge** as the filtration parameter changes. Various methods can be used in order to transform a discrete point set into a connected topological space. For example, simplicial complexes (see below) such as the Vietoris–Rips complex (VR complex) can be used, or a function can be defined over the domain of the data using an empirical distance function or a kernel density estimate (KDE) of the point cloud.

In this work, we use a VR complex to construct the filtration (discussed below). **As an illustration, suppose points are randomly sampled on three loops as displayed in Fig. 3a. The data points on their own do not form any loops (i.e., H_1 features), but from observing the data one may conjecture that that the underlying topological space has three loops. With persistent homology using the VR complex, each data point becomes the center of a ball with a radius t. The radius, t, is the parameter that determines a filtration as it increases from $t = 0$ to $t = \infty$. In Fig. 3a, the radius is $t = 0.5$ and some of the balls intersect and, therefore, become connected. As t increases, more balls inter-**

sect and eventually in the joined complex a loop forms as in Fig. 3b. The scale at which the loop forms and eventually gets filled in is tracked as the *birth* and *death* times of the features, and its persistence is the difference between the death and birth times (i.e., t-radius values at birth and death). A higher persistence means that feature survived longer in the filtration.

The birth and death times of all detected features are captured in a persistence diagram such as the one displayed in Fig. 3c. Features that last longer in the filtration are further from the diagonal line, y=x. If instead of the data used in Fig. 3 that were sampled on three circles, we considered data points randomly scattered in the same 2D window, there could still be H_1 features that form and die, but they may not persist as long in the filtration. In this section we provide more background on persistent homology including key concepts. ~~Next, we introduce some key components of persistent homology.~~

a. Simplicial complexes Simplicial complexes are the intermediate structures that are used to compute the homology of data on different scales. For our 3D halo data, the simplices we use are zero-simplices (vertices), one-simplices (edges), two-simplices (triangles), and three-simplices (tetrahedrons). More generally, a ~~A~~ geometric k-simplex is the convex hull of $k+1$ affinely independent points. A face of a simplex is another simplex obtained by removing zero or more points (e.g., a triangle has seven faces: itself, three edges, and the three vertices). An abstract simplicial complex, $\mathcal{K}$, is defined as a finite set of simplices such that (i) if $\sigma \in \mathcal{K}$, then every face of σ is also in $\mathcal{K}$, and (ii) if $\sigma_1, \sigma_2 \in \mathcal{K}$, then either $\sigma_1 \cap \sigma_2 \in \mathcal{K}$ or $\sigma_1 \cap \sigma_2 = \emptyset$. In this work, we use simplicial complexes to represent topological spaces, as they are the standard input to code to compute homology.

b. Filtrations ~~A filtration is a sequence of nested topological spaces.~~ To use persistent homology, the input is not just one simplicial complex, but a whole filtration or sequence of topological spaces (represented as simplicial complexes). In our setting, as the radius t increases, the simplicial complex grows to include more segments, triangles, and tetrahedra, but any simplicial complex that existed for a smaller value of t is also included in the larger simplicial complex. Consider the following more formal description. Given dataset $y_1, y_2, \ldots, y_n \in \mathcal{Y} \subseteq \mathbb{R}^3$, one common way to create a simplicial complex is to choose some $t \in \mathbb{R}$ such that $t \geq 0$ and replace each $y_i \in \mathcal{Y}$ with a ball of diameter t. The Vietoris–Rips complex at scale t (the t-VR complex) is created by representing each of these balls as a vertex, and creating a k-simplex anytime there are $k+1$ balls that pairwise intersect. Specifically:

$$VR_t(S) = \{\sigma \subseteq S \mid d(x,z) \leq t, \forall x, z \in \sigma, \} \qquad (2)$$

where $d(\cdot,\cdot)$ is the Euclidean distance [41, 42]. That is, $VR_t(S)$ is a simplicial complex containing the vertex set $\mathcal{S}$, edges between all the vertices that are separated by at most t, and triangles for sets of three vertices that have pairwise distances of at most t.

We obtain the VR filtration by increasing t from 0 to ∞ (recall that ~~here,~~ t is referred to as the *filtration parameter*).[5] Note that $VR_{t_1}(S)$ is a subset of $VR_{t_2}(S)$ (i.e., $VR_{t_1}(S) \subseteq VR_{t_2}(S)$) for $t_1 \leq t_2$. Sometimes, for the right selection of t and a dense enough sample, we can recover the underlying *true* homology of $\mathcal{Y}$ (see, e.g., Ref. [43]); however, using the whole sequence of complexes, we can recover information about $\mathcal{Y}$ with more relaxed sampling conditions.

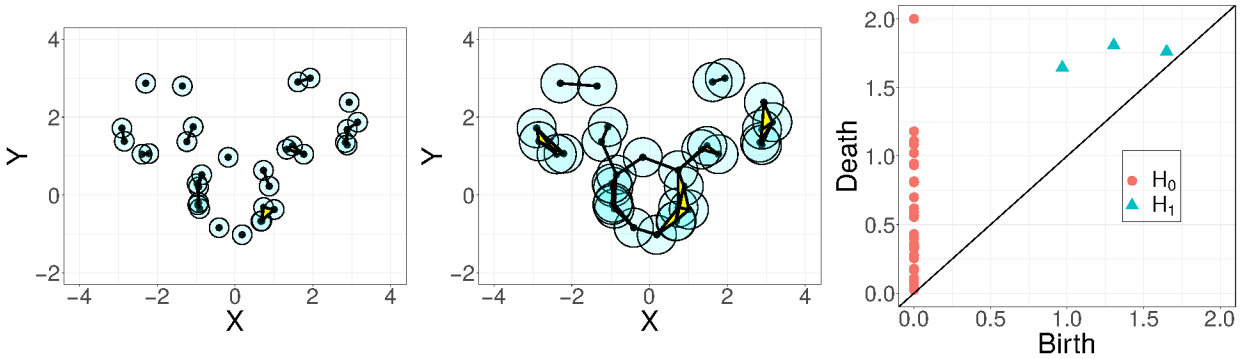
To derive the persistent homology for a VR filtration, the homology of $VR_t(S)$ is computed as t changes. If t is initialized at 0, then only the data points contribute to the homology. In our setting, this generally implies that at $t = 0$, each data point will represent an H_0 feature and no higher dimensional homological features exist yet. The evolving topological space is characterized by its homology as t increases toward ∞. For a $\mathcal{Y} \subseteq \mathbb{R}^3$, the persistent homology would then track the connected components (H_0), loops (H_1), and voids (H_2) that appear and disappear in the VR filtration. As was discussed previously, ~~A~~ an example of a VR filtration with a ~~2~~two-dimensional domain is presented in Fig. 3: Figures 3a and 3b display the data points with balls of diameter $t = 0.5$ and 1, respectively, along with the one- and two-simplices of the corresponding VR complex. Fig. 3c shows the persistence diagram for the data points using the VR filtration, which is discussed next.

c. Tracking Homology Generators The birth and death times of the homology group generators are displayed in a persistence diagram. These times correspond to values of the filtration parameter, which is the diameter of the balls t when considering a VR filtration. Suppose a filtration is defined over some data points $y_1, y_2, \ldots, y_n \in \mathcal{Y} \subseteq \mathbb{R}^3$, then a persistence diagram, $\mathbf{D}$, can be written as a multiset of points:

$$\mathbf{D} = \{(r_j, b_j, d_j) : j = 1, \ldots, |\mathbf{D}|\} \cup \Delta \qquad (3)$$

where (r_j, b_j, d_j) are the homology group dimension, the birth time, and the death time, respectively, of feature j, $|\mathbf{D}|$ indicates the number of homology group generators with $d_j > b_j$, and Δ represents a set of points on the diagonal (birth time = death time) with infinite multiplicity (which is included for mathematical reasons). The persistence diagram is a nice summary because small changes in the input data $\mathcal{Y}$ will result in only small changes in the diagram [44, 45], making the diagrams stable summaries of the data.

[5] In practice, the maximum filtration value t we consider correspond to the largest scales encompassed by a given galaxy/halo catalog.



(a) Data with $t = 0.5$ balls (b) Data with $t = 1$ balls (c) Persistence diagram

FIG. 3: Persistence diagram example where observations were sampled around three circles with noise. The data are displayed in (a) and (b) as black points (zero-simplices) with cyan balls with diameters of 0.5 and 1, respectively, along with the one- and two-simplices of the corresponding VR complexes. The persistence diagram for the VR filtration of the points is displayed in (c) with the the three circles indicated by the cyan triangles (H_1). The H_0 features represent the connected components, which all have birth times at 0.

Figures 3a and 3b show an example where the filtration parameter, t, increases from 0.5 to 1. In that interval, the homology changed from having 15 connected components (H_0**'s features**) and zero loops (H_1**'s features**) to having ~~5~~ **five** connected component and ~~1~~ **one** loop. The time in the filtration when homology features appear, the *birth* of the feature, and the time when a feature joins other features, the *death* of the feature, are captured in a persistence diagram. Fig. 3c displays the persistence diagram, where the location of each point represents the birth height (x-axis) and death height (y-axis) of a homological feature for a VR filtration, and the shape and color represent the homology group dimension. A point $(\cdot, x, x)$ on the diagonal represents a feature with a zero-length lifespan. The *persistence* of a point $(\cdot, b, d)$ is the length of the interval of the persistence parameter that supports that feature: $|d - b|$. In the persistence diagram, the distance from $(\cdot, b, d)$ to the diagonal is proportional to this value; in fact, ~~**the (Euclidean) distance to the diagonal is** $\frac{|d-b|}{\sqrt{2}}$~~ **the L_∞ distance of** $(\cdot, b, d)$ **is** $\frac{1}{2}|d - b|$**.** When working with empirical experimental or observational data**,** it is necessary to be concerned with the associated intrinsic noise of such measurements. This is especially important in the context of a spatial distribution of objects derived either from N-body simulations or galaxy catalogs. N-body simulations are limited by their spatial resolution, where their Monte Carlo sampling nature starts to breaks down and is overrun by the shot-noise. The astronomical observations are limited by imperfections including those related to involved instruments which contribute to measurements errors. For these reasons**,** it is important in a persistent homology analysis to be able to distinguish between real features really present in the target and those that are noise-induced transients. A notion of a topological significance can be derived in this context by considering features with longer lifetimes as more significant, and those with short lifetimes (i.e., closer to the diagonal) as topological noise [46]. Distinguishing between topological signal and noise is a problem of a great interest in real applications of TDA (e.g., Ref. [26]).

B. Persistence diagram summaries

While persistence diagrams and their individual features provide useful information about the topology of a data set, persistence diagrams are not easy objects to work with directly for statistical analyses. For example, the distance between two persistence diagrams can be calculated using metrics such as the bottleneck distance or the p-Wasserstein distance, but both are computationally expensive because they require finding a certain optimal matching between the features on each diagram; see Equation (B2) in the Appendix for the definition of the bottleneck distance. Fréchet means and medians have been defined for spaces of persistence diagrams [47], but are also computationally expensive and not necessarily unique (although ways around this exist as addressed in Ref. [48]). Instead, we consider transformations and summaries of persistence diagrams that make computations more tractable [31]. Below are several approaches that transform a persistence diagram into a functional summary, which are used in §IV to formulate test statistics for hypothesis tests.

a. Landscape Functions Landscape functions [33] are popular functional summaries of persistence diagrams [31, 49, 50], which are defined as follows. Let $\mathbf{D}_r = \{(b_j, d_j)\}_{j=1}^{n_r}$ be the finite set of off-diagonal points of a homology dimension r persistence diagram. Next, rotate the persistence diagram such that each point $(b_j, d_j) \in \mathbf{D}_r$ is mapped to $p_{r,j} = \left(\frac{b_j+d_j}{2}, \frac{d_j-b_j}{2}\right) \in \widetilde{\mathbf{D}}_r$. Isosceles right triangles are formed from each $p_{r,j}$ to the

base as

$$\Lambda_{p_{r,j}}(t) = \begin{cases} t - b_j & t \in [b_j, \frac{d_j+b_j}{2}] \\ d_j - t & t \in [\frac{d_j+b_j}{2}, d_j] \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

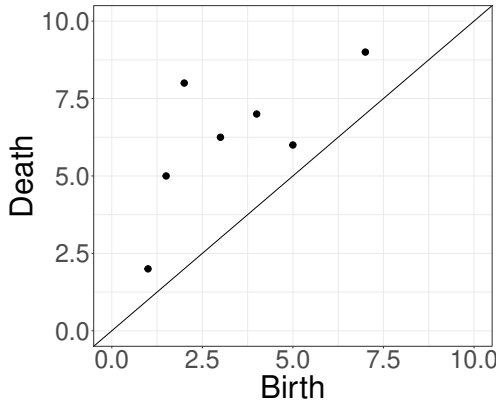
where $t \in [t_{\min}, t_{\max}]$. The persistence landscape is then defined as the following collection of functions

$$\lambda_{\mathbf{D}_r}(k,t) = \underset{p_{r,j} \in \widetilde{\mathbf{D}}_r,}{\text{kmax}} \Lambda_{p_{r,j}}(t), t \in [t_{\min}, t_{\max}], k = 1, \ldots, n_r, \tag{5}$$

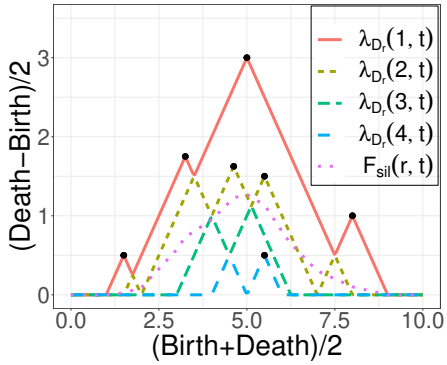
where kmax is the k-th largest value. An example of a persistence landscape function is displayed in Fig. 4. Rather than working with each k of $\lambda_{\mathbf{D}_r}(k,t)$ individually, a subset of the landscape layers can be concatenated to a long vector as

$$\mathcal{F}_{\text{land}}(\mathcal{I}, r, t) = \oplus_{k \in \mathcal{I}} \lambda_{\mathbf{D}_r}(k,t), \tag{6}$$

where $\mathcal{I}$ is the index set of the included landscape layers.



(a) Persistence Diagram



(b) Summaries

FIG. 4: A persistence diagram (a) along with its landscape functions and a weighted silhouette (b) for an arbitrary homology dimension r. The dotted pink curve is the weighted silhouette function with tuning parameter $p = 1$; the other four curves correspond to landscape functions $\lambda_{\mathbf{D}_r}(k,t)$ for $k = 1, \ldots, 4$.

b. Weighted Silhouette Functions Rather than working with each k of $\lambda_{\mathbf{D}_r}(k,t)$ from Equation (5) individually, weighted silhouette functions provide a way of combining the information in the collection of landscape functions. Silhouettes are weighted averages of the individual functions for homology dimension r defined as

$$\mathcal{F}_{\text{sil}}(r, t \mid p) = \frac{\sum_{j=1}^{n_r} |d_{r,j} - b_{r,j}|^p \Lambda_{p_{r,j}}(t)}{\sum_{j=1}^{n_r} |d_{r,j} - b_{r,j}|^p}, \tag{7}$$

where the $|d_{r,j} - b_{r,j}|^p$ act as weights that can give more emphasis or less emphasis to features with longer lifetimes depending on the user-specified parameter p. The form of these weights are suggested in Ref. [51]. An example of a weighted silhouette function is provided in Fig. 4b. More details and theoretical properties of landscapes and silhouettes can be found in Ref. [51].

c. **Betti and Euler Characteristic** *Functions* **The rth Betti number is the rank of the rth homology group (that is, the number of homological features of dimension r).** The Euler characteristic (EC) is a topological invariant and can be defined as the alternating sum of the **Betti numbers.** ~~**rank of the homology groups, where the rank of the rth homology group is the rth Betti number.**~~ As the persistent homology filtration parameter t changes and new features are born or old ones die, the Betti numbers and EC changes, allowing for the definition of Betti functions and an EC function. The Betti functions can be defined as

$$\mathcal{F}_{\text{betti}}(r,t) = |\{(r, b_j, d_j) : b_j \leq t, d_j > t\}|, \tag{8}$$

which indicates the number of dimension r homology group generators that persist in the filtration at time t. The only non-trivial homology groups for data in $\mathbb{R}^3$ are in dimensions 0, 1, and 2; thus, the Euler characteristic equation we use is

$$\mathcal{F}_{\text{ec}}(t) = \sum_{r=0}^{2} (-1)^r \mathcal{F}_{\text{betti}}(r,t). \tag{9}$$

Betti and EC functions have been used in applications [e.g., 23, 25, 52–56] and some of their theoretical properties have been explored [e.g., 35, 57–60]. **Before these recent uses of EC and Betti functions, the EC and a related concept *genus* were put forward as a new method for characterizing the topology of the Universe. Here, these statistics were used as a measure of the connectivity of the matter distribution in the Universe (e.g., Refs. [61–65]).**

There are a number of other summary functions of persistence diagrams that have been defined [e.g., 34, 66, 67]. For a general discussion of summary functions of persistence diagrams, including some theoretical properties, see Ref. [31].

IV. METHODS: TOPOLOGICAL HYPOTHESIS TESTS FOR LSS

A primary goal in this work is to develop a framework that can inferentially discriminate between different realizations of web-like geometric data structures such as the Cosmic Web. Our TDA-based framework allows for extracting information, encoded in large-scale galaxy/halo distribution, that goes beyond methods commonly used in cosmology N-point clustering statistics. The motivation is to detect differences between the DM halo spatial distributions (i.e., 3-manifolds) evolved in cosmological simulations where initial conditions were set to be either that of CDM or WDM-type. In this section we present a hypothesis testing framework using test statistics derived from the summaries of persistence diagrams presented in §III B. These topological hypothesis tests build on the work outlined in Ref. [31], including their notation.

The proposed hypothesis tests rely on permutation methods to compute the p-values. **We adopt this methodology because the distributions of the test statistics based on the functional summaries of the persistence diagrams are unknown. (The permutation tests are described below).** There ~~have been~~ **exist** some **results based on the** central limit theorem ~~results~~ for summary functions of persistence diagrams (e.g., Betti functions) and persistent homology-based hypothesis tests statistics using asymptotic theory (e.g., Refs. [35, 58, 68])~~, but they generally assume the data were drawn from a homogeneous Poisson point process.~~ **These results, however, generally assume the data were drawn from a homogenous Poisson point process.** Both the large and small-scale halo/galaxy distribution in the Universe cannot be described by a homogeneous Poisson process. Owing to the nature of initial conditions (i.e., an adiabatic Gaussian random field) and the gravitational instability (a mechanism responsible for the growth and evolution of the cosmic structures) halos spatial distribution is clustered with non-Gaussian features on small (non-linear) scales. Naturally, also the COCO DM simulations provide halo samples which are not close to resembling homogeneous Poisson point processes, which is discussed in Appendix §A. Furthermore, as we present below, the WDM and CDM samples are not independent of one another due to the cosmological simulation design: it is therefore necessary to use matched-pairs hypothesis tests.

A. Test statistic and p-value computations

For the proposed hypothesis tests, we consider two samples of observations,

$$\mathbf{Y}_1 = \{Y_{1,1}, \ldots, Y_{1,n_1}\}, \text{ and } \mathbf{Y}_2 = \{Y_{2,1}, \ldots, Y_{2,n_2}\} \tag{10}$$

where each $Y_{j,i}$, $i = 1, \ldots, n_j$ and $j = 1, 2$ is a data set of which a persistence diagram can be computed. For our cosmological simulation data, each $Y_{j,i}$ will have a set of points in $\mathbb{R}^3$, but, in general, the $Y_{j,i}$'s could take different forms; for example, each $Y_{j,i}$ could be an image of a fibrin network [31] or a brain artery tree [29]. **One sample represents the COCO-CDM, and the other sample represents the COCO-WDM data.**

Each observation from Equation (10) will have a corresponding persistence diagram

$$\mathbf{D}_1 = \{\mathbf{D}_{1,1}, \ldots, \mathbf{D}_{1,n_1}\}, \text{ and } \mathbf{D}_2 = \{\mathbf{D}_{2,1}, \ldots, \mathbf{D}_{2,n_2}\}. \tag{11}$$

These samples of diagrams can be used to test the hypotheses,

$$H_0 : \mathcal{P}_1 = \mathcal{P}_2 \text{ vs. } H_1 : \mathcal{P}_1 \neq \mathcal{P}_2, \tag{12}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the true underlying distributions of persistence diagrams from group 1 and 2, respectively.[6]

Given two samples of persistence diagrams, there are a number of possible ways to derive test statistics; we consider the functional versions of persistence diagrams presented in §III B as test statistics. These functional summaries can be understood as a map between the space of persistence diagrams, $\mathcal{P}$, to the space of functions, $\mathcal{F}$, defined as $\mathbb{F} : \mathcal{P} \longrightarrow \mathcal{F}$. Therefore, the diagrams from above can be used to define the collection of functional summaries with $j = 1, 2$ as

$$\mathbf{F}_j = \{F_{j,1} = \mathbb{F}(\mathbf{D}_{j,1}), \ldots, F_{j,n_j} = \mathbb{F}(\mathbf{D}_{j,n_j})\}. \tag{13}$$

A test statistic for the two-sample hypothesis test of Equation (12) can be derived using estimates of functional summaries. Letting $F_{j,i} = \mathbb{F}(\mathbf{D}_{j,i}), i = 1, \ldots, n_j$, and $j = 1, 2$ (see Equation (13)), the mean functional summaries are defined as

$$\bar{F}_j(t) = n_j^{-1} \sum_{i=1}^{n_j} F_{j,i}(t). \tag{14}$$

Then our test statistic for the different functional summaries is based on the following distance between mean functional summaries,

$$d(\bar{F}_1, \bar{F}_2) = \int_{\mathbb{T}} |\bar{F}_1(t) - \bar{F}_2(t)| dt, \tag{15}$$

where $\mathbb{T}$ defines the domain of the functions. Note that the $\mathbb{T}$ is related to the range of values of the filtration parameter, which depends on the functional summary. For example, for the Euler characteristic function, the $\mathbb{T}$ covers the range of the filtration parameter, but for the landscape and silhouette functions it represents the range of a transformed filtration parameter since the persistence diagram is rotated.

[6] Probability measures can be theoretically defined on a space of persistence diagrams (with a Wasserstein metric) as presented in Ref. [69].

a. Matched-Pairs Permutation Test. Since the distributions of the test statistics of Equation (15) for the different functional summaries we consider are unknown, the p-values for the two-sample hypothesis tests can be computed using the usual permutation testing framework. The general procedure is to randomly assign the $n_1 + n_2$ functional summaries into two groups, because this random assignment is consistent with the null hypothesis (H_0) from Equation (12) where the two groups follow the same distribution. **In other words, because the null hypothesis states that the two samples were drawn from the same distribution of persistence diagrams, when we take the null hypothesis as true it should not matter to which group, 1 or 2, that a given sample is assigned. We can then create a number of random group assignment permutations. For each ~~Using the~~ random group assignments~~, the~~, a** new mean functional ~~**summaries are**~~ **summary is** estimated **for each group** using Equation (14), $\widetilde{F}_1^{(l)}$ and $\widetilde{F}_2^{(l)}$, which are used to compute the distance $d(\widetilde{F}_1^{(l)}, \widetilde{F}_2^{(l)})$ from Equation (15), for $l = 1, \ldots, n_l$ random permutations [31]. The resulting (approximate) permutation p-value can then be computed as

$$p_{\text{perm}} = n_l^{-1} \sum_{l=1}^{n_l} I(d(\widetilde{F}_1^{(l)}, \widetilde{F}_2^{(l)}) \geq d(\bar{F}_1, \bar{F}_2)), \qquad (16)$$

where $I(A)$ is an indicator function that takes the value 1 if A is true and 0 if A is false.

In summary, a standard permutation p-value is computed by assuming the null hypothesis is true (i.e., assuming no difference between the distributions of the two samples of persistence diagrams), and randomly permuting the group assignments a number of times. For each random permutation, a test statistic is computed from Equation (15). The collection of all these test statistics provides an approximation of the distribution of the test statistic when the null hypothesis is true. Then a p-value is computed by comparing the observed test statistic (i.e., the test statistic using the original groupings for the two samples), and estimating the probability of observing the test statistic that we did or one further in the tail of the null distribution. As usual, a small p-value is evidence *against* the null hypothesis.

When the two sets of samples are independent, the above permutation p-value is reasonable. However, as explained in §II, the CDM and WDM COCO data, $\mathbb{Y}_c$ and $\mathbb{Y}_w$, are not independent due to the initial conditions of the simulations. Instead, the samples $\mathbf{Y}_{c,i}$ and $\mathbf{Y}_{w,i}$ for $i = 1, \ldots, 77$ have a similar spatial structure which should be accounted for in the computation of the permutation p-values. Therefore, we consider a matched-pairs version of the permutation p-values. **The matched-pairs approach to hypothesis testing is common in settings where there is a clear connection between individual samples in two groups (e.g., studies carried out on monozygotic twins, pre- and post-assessments on participants that were randomly assigned treatments and each participant is matched with themselves). In our setting with CDM and WDM data, the matching is necessary because of the same initial conditions used in each run of the COCO simulation.**

The difference between this matched-pairs version and the permutation test outlined above is in how the two groups are randomly assigned for each permutation. The matched-pairs permutation involves randomly selecting one of the two matched samples to go into each of the two groups (e.g., one of $\mathbf{Y}_{c,i}$ or $\mathbf{Y}_{w,i}$ will be randomly assigned to group 1, and the other will be assigned to group 2). The matched-pairs permutation p-value is then defined in the same manner as above, as

$$p_{\text{matched}} = \sum_{l=1}^{n_l} I(d(\widetilde{F}_{1,\text{matched}}^{(l)}, \widetilde{F}_{2,\text{matched}}^{(l)}) \geq d(\bar{F}_1, \bar{F}_2)), \qquad (17)$$

where $\widetilde{F}_{j,\text{matched}}^{(l)}, j = 1, 2$, are the mean functional summary for permutation l using the matched-pairs random assignment. This matched-pairs permutations p-value computation accounts for correlations between the COCO CDM and WDM samples by including one of the two matched samples within each group for each permutation, but randomizing which label (CDM or WDM) is assigned.

V. INVESTIGATION OF COCO SIMULATION DATA

In order to investigate differences between the CDM and WDM COCO samples of MW-analog halo neighborhoods described in §II B, $\mathbb{Y}_c$ and $\mathbb{Y}_w$, respectively, we carry out the two-sample hypothesis tests defined in Equation (12) and described in the previous section. The test statistics are based on the functional summaries of persistence diagrams outlined in §III B, along with several other methods discussed below. The comparison methods include a test statistic that uses persistence diagrams directly (rather than a functional summary of them) and non-TDA functional summaries that capture second-order properties of spatial point processes. The collective goals of the test statistics considered are (i) to detect differences between CDM and WDM MW-analog halo neighborhoods, and (ii) to understand and interpret any detected differences (e.g., the distance scale at which differences occur).

In addition to the test statistics using the functional summaries presented in §IV, we also consider other approaches. One method is the persistence diagram-based test (PDT) of Ref. [32] which has a test statistic defined using distances between persistence diagrams. We also consider two functional summaries of spatial point processes which do not use persistence diagrams, namely

the G-function and the two-point correlation function (2PCF). The G-function gives the distribution function of the nearest-neighbor distances, and 2PCF uses the Landy-Szalay estimator [70] and is one of the most basic and fundamental objects used to study clustering in cosmology [71]. These methods are described in more detail in Appendix §B. The p-values for these additional tests are also carried out using permutations, and we have also adapted them to work for our matched-pairs design.[7]

A. Hypothesis testing results

All the hypothesis tests that use statistics derived from persistence diagrams (including the PDT) use the same persistence diagrams, which were computed using a VR filtration. These computations were carried out with *Ripser* [72]. P-values were computed for the test statistics discussed previously based on 20,000 permutations using the traditional and matched-pairs permutation methods of §IV A. The results are displayed in Table I. Below we discuss the resulting p-values, and in the next section we investigate and interpret where the differences are most pronounced.

Overall, statistically significant differences with p-values < 0.001 are apparent between the CDM and WDM MW-analog DM halo neighborhoods samples, and $p_{\text{perm}} \geq p_{\text{matched}}$ for all test statistics considered.[8] Since the two sets of samples are from different populations (CDM vs. WDM COCO data), it is a positive result that our proposed tests are able to detect differences. For H_0 and H_1, all the function-based tests had $p_{\text{matched}} < 0.001$, and this was also the case for the G-function and 2PCF test statistics. PDT has p_{perm} and $p_{\text{matched}} \leq 0.003$ for H_0, but higher p-values for H_1 with $p_{\text{perm}} = 0.255$ and $p_{\text{matched}} = 0.036$. Because the PDT uses the bottleneck distance, only one H_1 feature on each of the persistence diagrams contribute to the test statistic for each MW-analog halo neighborhood sample, while the functional summary-based test statistics considers all the features on the persistence diagrams (except for the landscape functions which only includes features that contribute to the first 10 layers).

Aside from the silhouette function tests, the H_2 p-values are < 0.01 for both p_{perm} and p_{matched}. The p_{perm} for the silhouette function tests are > 0.10, but then drop below 0.01 for p_{matched}. The EC function test statistic, similar to the related Betti function test statistics, has both p_{perm} and $p_{\text{matched}} \leq 0.001$. For the TDA-based test statistics, the EC, Betti, and Landscape function test statistics appear to be best able to detect differences between the CDM and WDM MW-analog halo neighborhood samples for both the traditional and matched-pairs permutation tests across the three homology dimensions (H_0, H_1, H_2). Tuning could be carried out for the landscape function tests to find which landscape function layers are most informative at detecting differences. Since the results with the first 10 layers performed well, we did not consider tuning for this analysis.

Given that we only have one COCO-CDM and one COCO-WDM realization, and that we seek to evaluate the performance of the proposed test statistics when the null hypothesis is true, we consider bootstrap samples of the data from the CDM data and from the WDM data. The distribution of the p-values when the null hypothesis is true should follow a uniform distribution. Details of this simulation study and the results are presented in Appendix §C. Overall, we find the p-values resulting from proposed test statistics based on the functional summaries of persistence diagrams under the null hypothesis are generally consistent with uniform distributions.

B. Interpretation of results

In this section, we explore the Betti functions in more detail and develop other visualizations to aid in the interpretation of the results in order to investigate the scales at which the differences between the CDM and WDM MW-analog halo neighborhood samples occur and are significant. Since our interest is in where the test statistics diverge, the mean difference function is displayed where the signal is based on the matched data in the CDM and WDM COCO MW-analog halo neighborhood samples using

$$\bar{F}_{\text{diff}}(t) = n_s^{-1} \sum_{i=1}^{n_s} (F_{c,i}(t) - F_{w,i}(t)) \tag{18}$$

where $F_{c,i}(t)$ and $F_{w,i}(t)$ are functional summaries for CDM and WDM sample i, respectively, and $n_s = 77$. Additionally, 95% global confidence bands are computed using the bootstrap approach outlined in Section 3.2 of Ref. [31], with 1000 bootstrap samples.[9]

The CDM and WDM MW-analog halo neighborhood samples' persistence diagrams were generated using using a VR filtration. For example, Fig. 5 displays the persistence diagrams for the COCO CDM and WDM samples of Fig. 2a and 2b, respectively. The CDM and WDM persistence diagrams in this example share a similar pattern where generally the H_0 features are all connected by around a filtration parameter value of 1, H_1 features persist longer than the H_2 features across the range of

[7] **Data and code associated with this work is available at** `https://github.com/JessiCisewskiKehe/DarkMatterTDA`.

[8] If our test statistics were Gaussian distributed, a p-values $<$ 0.001 would correspond to $> 3\sigma$ significance.

[9] Note that the hypothesis tests use L_1 distances between functions (see Equation (15)) while the confidence bands are investigating differences across the functions.

TABLE I: COCO data results. Permutation p-values (p_{perm}) and matched permutation p-values ($p_{matched}$) for tests comparing the CDM and WDM MW-analog halo neighborhood samples. The p-values are rounded to three decimal places and are based on 20,000 permutations as described in §IV A.

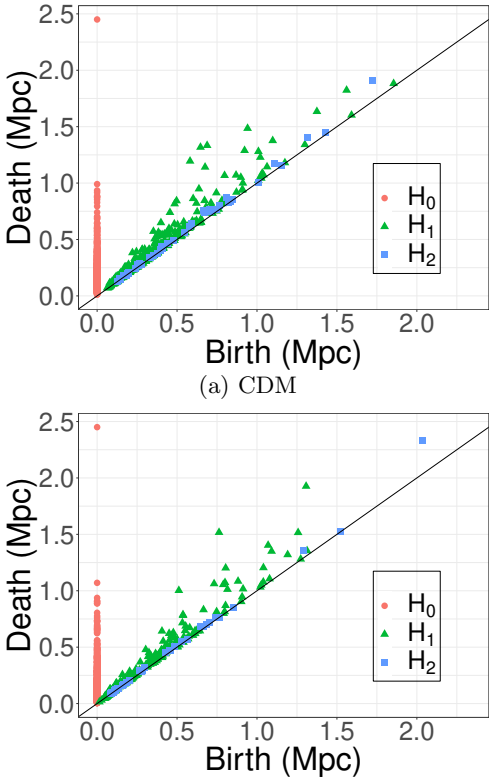
Test statistic	Notation	Homology dimension	p_{perm}	p_{matched}
Landscape	$\mathcal{F}_{\text{land}}(1:10,0,t)$	0	0	0
Silhouette	$\mathcal{F}_{\text{sil}}(0,t \mid p=0.5)$	0	0.009	0
Silhouette	$\mathcal{F}_{\text{sil}}(0,t \mid p=1)$	0	0.001	0
Silhouette	$\mathcal{F}_{\text{sil}}(0,t \mid p=2)$	0	0	0
Betti	$\mathcal{F}_{\text{betti}}(0,t)$	0	0.001	0
PDT	$\mathcal{T}_{\text{PDT}}(D_{1,\cdot\mid 0}, D_{2,\cdot\mid 0} \mid \infty, 1)$	0	0.003	0
Landscape	$\mathcal{F}_{\text{land}}(1:10,1,t)$	1	0	0
Silhouette	$\mathcal{F}_{\text{sil}}(1,t \mid p=0.5)$	1	0.009	0
Silhouette	$\mathcal{F}_{\text{sil}}(1,t \mid p=1)$	1	0.007	0
Silhouette	$\mathcal{F}_{\text{sil}}(1,t \mid p=2)$	1	0.014	0
Betti	$\mathcal{F}_{\text{betti}}(1,t)$	1	0	0
PDT	$\mathcal{T}_{\text{PDT}}(D_{1,\cdot\mid 1}, D_{2,\cdot\mid 1} \mid \infty, 1)$	1	0.255	0.036
Landscape	$\mathcal{F}_{\text{land}}(1:10,2,t)$	2	0	0
Silhouette	$\mathcal{F}_{\text{sil}}(2,t \mid p=0.5)$	2	0.123	0.003
Silhouette	$\mathcal{F}_{\text{sil}}(2,t \mid p=1)$	2	0.158	0.009
Silhouette	$\mathcal{F}_{\text{sil}}(2,t \mid p=2)$	2	0.135	0.008
Betti	$\mathcal{F}_{\text{betti}}(2,t)$	2	0.001	0
PDT	$\mathcal{T}_{\text{PDT}}(D_{1,\cdot\mid 2}, D_{2,\cdot\mid 2} \mid \infty, 1)$	2	0.009	0
Euler characteristic	$\mathcal{F}_{\text{ec}}(t)$	0-2	0.001	0
G-function	$\mathcal{F}_{\text{G}}(t)$	N/A	0	0
2PCF	$\mathcal{F}_{\text{2PCF}}(t)$	N/A	0	0

birth times. The H_0 feature plotted on both diagrams at (0, 2.57) represents an H_0 feature that in fact persists indefinitely and should, technically, be plotted at a death time of infinity.

The mean differences (WDM - CDM) of the Betti functional summaries are displayed in Fig. 6 along with the corresponding 95% confidence bands. Overall, these summaries suggest that the CDM and WDM samples differ on shorter distance scales, but then start to resemble each other at longer distance scales in keeping with Fig. 1. Recall that the Betti functions count the number of features that are persistent at the filtration parameter values (i.e., the x-axis) so by considering the average difference of the Betti functions we observe at which scales the number of features differ between the CDM and WDM. For H_0, the number of features, on average, for the CDM data is larger than the number for the WDM for distances until scales of around 0.4 Mpc, and then the number of WDM features is slightly higher than the number of CDM features until distances of ∼0.75 Mpc. The number of H_1 features is greater, on average, for the WDM data over the CDM data when $t \leq 0.13$ Mpc, and then the CDM has more H_1 features until around 0.9 Mpc. A similar pattern is observed with the H_2, but the average differences between the CDM and WDM are within only two H_0 features.

While Betti functions capture the number of features that persist across the filtration parameter values, we defined analogous functions that instead capture the maximum persistence (MaxPers) and average persistence (AvePers), which are displayed in Fig. 7a and Fig. 7b, respectively. Similar to the plots in Fig. 6, the mean difference (CDM-WDM) of these MaxPers and AvePers functions for the matched samples were computed. However, for Fig. 7, in order to visualize the variability in the mean differences, pointwise error bars (± one standard error) are included. The filtration parameter grid ranges from 0 to 2.5 Mpc with a spacing of 0.05. This is a lower resolution than the Betti function figures, which we adopt here in order to be able to improve the visibility of the individual error bars. There are larger differences between CDM and WDM MaxPers in H_0, H_1, and H_2 for $t \lesssim 1.85$ Mpc: generally the H_0 MaxPers are greater for WDM than CDM, the H_1 MaxPers is greater for CDM than WDM at scales $\lesssim$1.1 Mpc when this tendency switches and WDM has greater MaxPers, and the H_2 MaxPers are higher for CDM than WDM. A similar pattern is apparent with the AvePers functions except the H_1 AvePers are similar for CDM and WDM until scales around 1 Mpc, after which WDM generally has greater AvePers until around 2 Mpc.

Basic spatial point process summary functions, such as the 2PCF, are commonly employed tools in cosmological large-scale structure study. To quantify the degree to which our persistence diagrams provide new information over these standard statistics, we calculate and show



(a) CDM

(b) WDM

FIG. 5: Persistence diagrams for the MW-analog halo neighborhood sample for (a) the CDM data of Fig. 2a, and (b) the WDM data of Fig. 2b.

mean difference functions for the G-functions and 2PCFs in Fig. 8a and Fig. 8b, respectively. The plotted data indicate that the WDM functions take, on average, greater values than the CDM variants for $t \lesssim 0.5$ Mpc. This result points toward a similar direction as what we observed with the differences in the H_0 Betti function mean differences displayed in Fig. 6a. This is not surprising since the H_0 Betti functions, the G-functions, and the 2PCFs have different ways of assessing the closeness of the halos within the samples. However, the H_1 and H_2 Betti functions, together with the MaxPers and AvePers methods, appear to detect differences between the CDM and WDM at different scales, suggesting that they provide distinct information from the spatial point process functions. In particular, the H_1 and H_2 functions suggest that as the halos become connected (i.e., the death of H_0 features), the CDM and WDM models are forming loops and voids (i.e., H_1 and H_2 features, respectively) in different ways. Also, the MaxPers and AvePers of the H_0 features differ between the CDM and WDM data on different scales than those of the H_0 Betti functions.

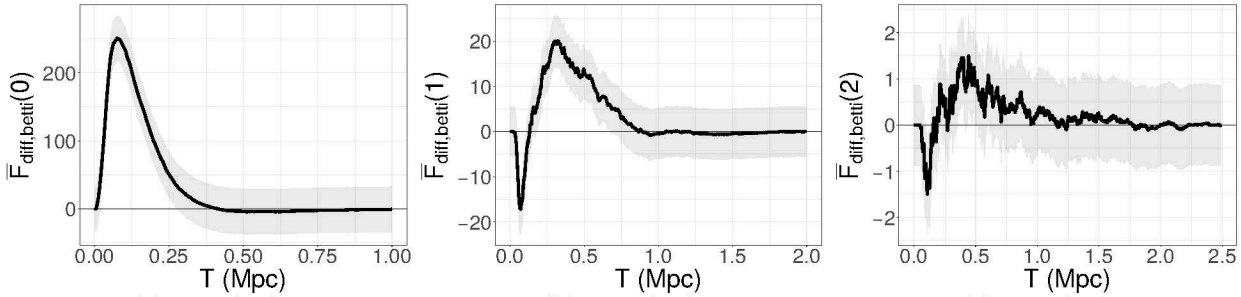
VI. CONCLUSION

The LSS contains valuable information about the composition, and evolution, and the physical nature of the Universe. TDA tools such as persistent homology provide a novel opportunity to extract this information from cosmological data. While TDA-based approaches have been applied in various fields of statistical studies, its application to cosmological data and analysis is still in its infancy. In this paper, we introduced a hypothesis testing framework built on persistent homology that extends the work of Ref. [31] in order to compare topological summaries of MW-analog halo neighborhoods (3 Mpc spheres) evolved under two different DM models: CDM and WDM (WDM thermal relic mass: 3.3 keV). Next, we have assessed the sensitivity and robustness of this framework in the context of differentiating between CDM and WDM variants. The proposed collection of test statistics based on persistence diagrams uses summaries that were recently proposed in the literature [23, 31, 33, 49–51], and are easier to work with than the original persistence diagrams. The results of the persistence diagram-based functional summaries were compared to two spatial point process functional summaries (G-functions and 2PCF) and test statistics that use persistence diagrams directly (PDT) [32].

We showed empirically that such a framework is able to infer differences between CDM and WDM, and investigated the scales at which differences occur. While most of the test statistics were able to detect statistically significant differences with $p_{\text{matched}} \leq 0.009$ for all tests considered except the PDT for H_1 (§V A, especially Table I), the persistent homology-based functional summaries appear to detect differences between the CDM and WDM data on different scales from the spatial point process functional summaries (§V B).

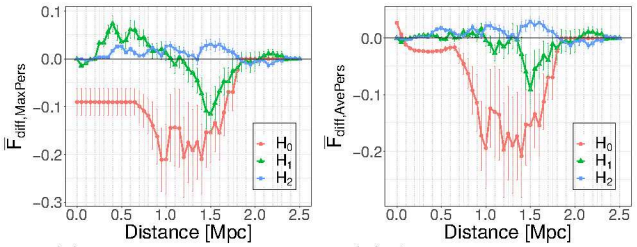
Our results imply that the homology properties of clustered CDM and WDM haloes distributions are very different on small scales. CDM haloes are distributed across a larger number of clusters (homology dimension 0) than WDM haloes, especially at the ~ 80 kpc filtration scale (Fig. 6a), although the clusters that form in WDM are more persistent on average (Fig. 7b). The 80 kpc scale is also where the two process functions return the biggest difference between the models—in both cases an excess of clustering in CDM relative to WDM—plus the filtration scale at which WDM features more loops (homology dimension 1) than CDM. We thus build a picture in which CDM rapidly builds up a large number of small clusters, whereas WDM builds a smaller number of clusters, many of which will be rapidly converted into loops.

This picture is consistent with the formation of haloes in and around cosmological filaments. In CDM, the distribution of filaments extends to near arbitrarily small scales and fills much of configuration space, whereas the WDM cutoff restricts WDM haloes to lie along large filaments and so their spatial distribution is much more constrained. Therefore, the dispersed CDM haloes form



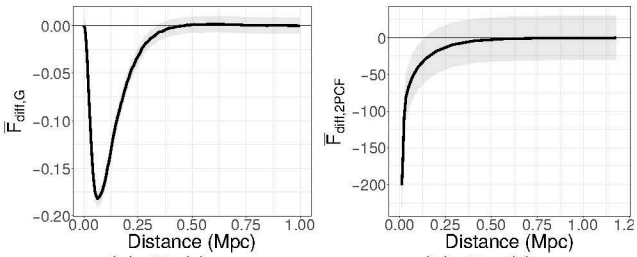
(a) $\mathcal{F}_{\text{betti}}(0,t)$ (b) $\mathcal{F}_{\text{betti}}(1,t)$ (c) $\mathcal{F}_{\text{betti}}(2,t)$

FIG. 6: Mean differences (CDM-WDM) of the Betti functional summaries along with 95% confidence bands (shaded regions) for the noted functional summaries. The x-axis limits were set to highlight the non-zero mean differences regions.



(a) Max persistence (b) Average persistence

FIG. 7: Mean differences (CDM-WDM) of the maximum (a) and average (b) persistences ± one standard error. Means and standard errors were computed every 0.05 Mpc between 0 and 2.5. Gray dotted vertical lines are plotted every 0.10 Mpc. Note that the H_0 feature that persists indefinitely has been removed from this analysis.



(a) $\mathcal{F}_{\text{G}}(t)$ (b) $\mathcal{F}_{\text{pcf}}(t)$

FIG. 8: Mean differences (CDM-WDM) of the point process functions along with 95% confidence bands (shaded regions).

large numbers of small, isolated clusters, whereas WDM haloes are quickly joined up along cosmological filaments into loops.

The question remains as to whether this difference between the models can be detected in the spatial distribution of observed Local Group galaxies. One will have to select haloes that are likely to form a galaxy, where most of the haloes that we included in this study will be below the HI cooling limit and thus dark [73, 74]. Reducing the number of haloes available in this manner will likely lead to a reduction in the statistical significance of differences between the models' persistent homology properties: it is therefore imperative to make halo selections based on, for example, peak halo mass or a semi-analytic models [75] to confirm the potential persistent homology has for understanding which DM models best describe the Local Group.

Appendix A: Distributional Assumptions of DM Samples

In this section, we carry out tests to show that the spatial distributions of the halo samples do not follow a homogeneous Poisson point process (i.e., *complete spatial randomness*, CSR), which then precludes the use of many theoretical results that rely on that assumption; see §IV for a brief discussion about some asymptotic results in persistent homology. Ref. [76] describe a straightforward Monte Carlo test for checking CSR. Using the same number of observations (i.e., the number of halos in the MW-analog halo neighborhoods) and the same window volume (i.e., a sphere with radius 3 Mpc), N_{MC} Monte Carlo realizations are generated assuming CSR and then their G-functions are estimated; see Appendix §B and Equation (B4) for background on G-functions. Then a global envelope is defined using the N_{MC} summary functions based on the maximum absolute deviation of the simulated summary functions from the (known) theoretical summary function (assuming CSR). If the summary function for the observations are outside the band, then that is evidence against CSR for those data.

The global envelope was computed by generating N_{MC} realizations of a homogeneous Poisson process within a sphere of radius 3 Mpc using rejection sampling. The number of points was set to match the number of halos in each of the MW-analog halo neighborhood samples. Then a G-function was estimated for each sample using the *G3est* function in the *spatstat* R package. For each simulated G-function, the maximum absolute deviation was computed using the true G-function of the corresponding Poisson process, defined as

$$\mathcal{F}_G(t) = 1 - e^{-\frac{4}{3}\pi\hat{\lambda}t^3} \tag{A1}$$

where $\hat{\lambda}$ is the intensity estimated as the number of points divided by the volume of the sphere. The interpretation is that if the observed G-function is outside the envelope for any value t Mpc, then we can reject CSR at a significance level of $1/(1 + N_{\text{MC}})$ (Ch. 10, Ref. [76]). Using $N_{\text{MC}} = 19$, these global envelopes were computed for all 77 samples of the CDM and WDM data, and all observed G-functions have regions outside the envelopes. For illustration purposes, the resulting global envelope and observed G-functions for the CDM and WDM MW-analog halo neighborhood samples from Fig. 2 are displayed below in Fig. 9. **Notice that the estimated G-functions for the CDM and WDM samples are outside the gray band for scales below 0.375 Mpc suggesting that they are not consistent with CSR.**

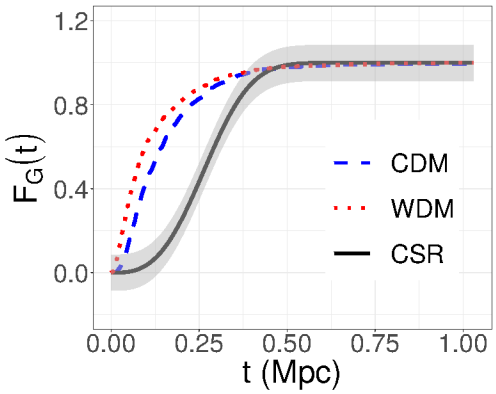


FIG. 9: The estimated spatial point process summary functions for the CDM (dashed blue line) and WDM (dotted red line) MW-analog halo neighborhood samples displayed in Fig. 2, along with the theoretical summary function assuming CSR (solid black line) and its global envelope (gray region) using $N_{MC} = 19$ samples. Because this is a global envelope, we can reject the hypothesis that the CDM and WDM samples were generated from homogenous Poisson point process at the $1/(1 + N_{MC}) = 0.05$ level of significance.

Appendix B: Comparison Methods in COCO Analysis

In addition to the test statistics proposed based on functional summaries of persistence diagrams, we include three comparison test statistics in our investigation of the COCO simulation data presented in §V. The comparison methods are the Persistence Diagram Test (PDT), and test statistics derived using the G-function and two-point correlation function (2PCF) which are popular functional summaries of spatial point processes. The comparison methods are described below.

a. Persistence Diagram Test (PDT) Ref. [32] developed a two-sample test that compares persistence diagrams rather than functional summaries of persistence diagrams. The PDT test statistic takes the following form,

$$\mathcal{T}_{\text{PDT}}(D_{1,1:n_1|r}, D_{2,1:n_2|r} \mid p, q) = \sum_{l=1}^{2} \frac{1}{2n_l(n_l-1)} \sum_{i=1}^{n_l}\sum_{j=1}^{n_l} W_p(D_{l,i|r}, D_{l,j|r})^q \quad \text{(B1)}$$

where $D_{l,1:n_l|r}$ is a set of n_l persistence diagrams for homology dimension r from population $l = 1, 2$, q satisfies $1 \le q < \infty$, and $W_p(\cdot,\cdot)$ is the p-Wasserstein distance, with $1 \le p \le \infty$. In this work, we set $q = 1$ and $p = \infty$. The W_∞ distance is also known as the *bottleneck distance*, and is defined as

$$W_\infty(D_1, D_2) = \inf_{\eta: D_1 \to D_2} \sup_{x \in D_1} \|x - \eta(x)\|_\infty \quad \text{(B2)}$$

where D_1 and D_2 are persistence diagrams, η defines a bijection between the two persistence diagrams that allows for matches to the diagonal Δ, and $\|\cdot\|_\infty$ is the L_∞ norm in $\mathbb{R}^2$ computed between the birth and death coordinates of x and $\eta(x)$ for a fixed homology dimension r.

b. Spatial Point Process Functions In order to investigate properties of the spatial distributions of the data, which in our setting is the location of the DM halos, we consider a popular functional summary of spatial point processes, the G-function[10], along with the 2PCF which is commonly used in cosmology research. The G-function is defined below and estimated using the implementation for three-dimensional point patterns in the R package `spatstat` [77, 78];[11] see Ref. [76] for more details. The 2PCF uses the Landy-Szalay estimator [70]

The G-function and 2PCF are estimated for each sample of the DM halos, $\mathbf{Y}_{k,i} \in \mathbb{R}^{n_i \times 3}$ for $k = w, c$ and $i = 1, \ldots, 77$. Given a sample $\mathbf{Y} \in \mathbb{R}^{n \times 3}$, let each point be denoted by $Y_i = (Y_{i,1}, Y_{i,2}, Y_{i,3})$ for $i = 1, \ldots, n$. Define a distance function, ρ, as

$$\rho(x, \mathbf{A}) = \inf\{\|x - a\| : a \in \mathbf{A}\} \quad \text{(B3)}$$

which represents the shortest distance between some point $x \in \mathbb{R}^3$ and a closed set $A \subset \mathbb{R}^3$. The G-function gives the distribution function of the nearest neighbor distances, and can be defined as

$$\mathcal{F}_G(t) = \mathbb{P}(\rho(Y_i, \mathbf{Y}_{-Y_i}) \le t \mid Y_i \in \mathbf{Y}) \quad \text{(B4)}$$

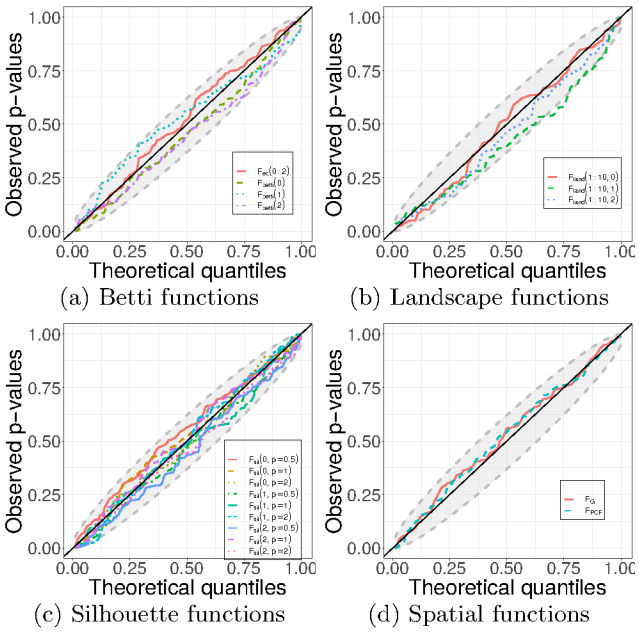
where $\mathbf{Y}_{-Y_i}$ is the set of points $\mathbf{Y}$ excluding the point Y_i. The Kaplan-Meier estimator of Ref. [79] is used to address the edge effects (i.e., boundary issues).

[10] Also referred to as the "nearest-neighbor distance distribution function"

[11] The R function from the `spatstat` package is `G3est`.

Appendix C: Distribution of p-Values Under the Null Hypothesis

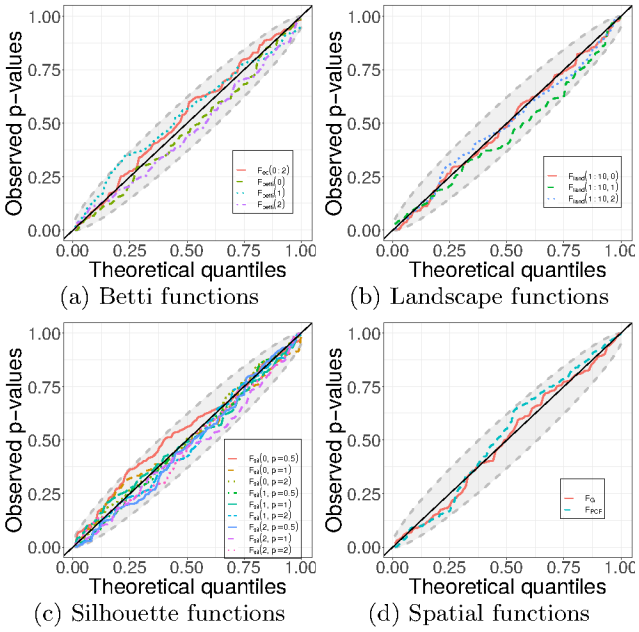
The results of the proposed hypothesis tests are presented in §V A. Many of the test statistics find statistically significant differences between the CDM and WDM models with p-values ≤ 0.001. In order to verify that the test statistics do not inappropriately reject the null hypothesis when the null hypothesis is true (i.e., when both groups come from either CDM or WDM), we carry out the following experiment. We repeatedly generate two sets of boostrap realizations from either the CDM or WDM samples, and then compute permutation p-values for the test statistics presented in the main text. The distribution of the p-values in this setting where the null hypothesis is true should follow a uniform distribution. To compute one p-value, two bootstrap samples (with replacement) of 77 MW-analog halo neighborhoods are selected from the CDM (WDM) data. Then the hypothesis testing framework presented in §IV A is used to compute a traditional permutation p-value (since the matched pairs design is not present in this setting) using 20,000 permutations. This computation is repeated for 100 independent iterations for the CDM (WDM) data with the same sampled indexes used for the CDM and WDM bootstrap samples. Fig. 10 and Fig. 11 display the results for the CDM and WDM samples, respectively, as uniform quantile-quantile plots with 99% pointwise bands based on the distribution of order statistics of uniform random variables (i.e., Beta(k, $n+1-k$) where k is the order and $n = 100$). The resulting p-values for each test statistic are generally consistent with uniform distributions. The CDM H_1 landscape function p-values (Fig. 10b) have some values that are not within the 99% confidence band, but this does not occur with the WDM H_1 landscape function p-values (Fig. 11b) nor with the other landscape function p-values so it appears to not be a reason for concern about the landscape function-based test statistics.



(a) Betti functions (b) Landscape functions

(c) Silhouette functions (d) Spatial functions

FIG. 10: Uniform quantile-quantile plots of the 100 permutations p-values calculated for each test statistic using bootstrap realizations of the CDM MW-analog halo samples. Each bootstrap sample includes 77 MW-analog halo neighborhoods, and 20,000 permutations were used to compute each p-value.



(a) Betti functions (b) Landscape functions

(c) Silhouette functions (d) Spatial functions

FIG. 11: Uniform quantile-quantile plots of the 100 permutations p-values calculated for each test statistic using bootstrap realizations of the WDM MW-analog halo samples. Each bootstrap sample includes 77 MW-analog halo neighborhoods, and 20,000 permutations were used to compute each p-value.

ACKNOWLEDGMENTS

This research was performed using the compute resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) 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 Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. This work used the DiRAC Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). This equipment was funded by BIS National E-infrastructure capital grant

ST/K00042X/1, STFC capital grants ST/H008519/1 and ST/K00087X/1, STFC DiRAC Operations grant ST/K003267/1 and Durham University. DiRAC is part of the National E-Infrastructure. This project has also benefited from numerical computations performed at the Interdisciplinary Center for Mathematical and Computational Modeling (ICM) University of Warsaw under grants #no GB79-7, GA67-17 and G63-3. JCK and BTF acknowledge support from NSF under Grant Numbers DMS 2038556 and 1854336. WAH and PD acknowledge the support from the Polish National Science Center within research projects no. 2018/31/G/ST9/03388, 2020/39/B/ST9/03494. MRL acknowledges support by a Grant of Excellence from the Icelandic Research Fund (grant number 206930).

[1] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. White, The evolution of large-scale structure in a universe dominated by cold dark matter, The Astrophysical Journal **292**, 371 (1985).

[2] P. Bull, Y. Akrami, J. Adamek, T. Baker, E. Bellini, J. B. Jimenez, E. Bentivegna, S. Camera, S. Clesse, J. H. Davis, *et al.*, Beyond ΛCDM: Problems, solutions, and the road ahead, Physics of the Dark Universe **12**, 56 (2016).

[3] J. S. Bullock and M. Boylan-Kolchin, Small-scale challenges to the ΛCDM paradigm, Annual Review of Astronomy and Astrophysics **55**, 343 (2017).

[4] L. Perivolaropoulos and F. Skara, Challenges for ΛCDM: An update, arXiv preprint arXiv:2105.05208 (2021).

[5] R. van de Weygaert, G. Vegter, H. Edelsbrunner, B. J. T. Jones, P. Pranav, C. Park, W. A. Hellwing, B. Eldering, N. Kruithof, E. G. P. P. Bos, J. Hidding, J. Feldbrugge, E. ten Have, M. van Engelen, M. Caroli, and M. Teillaud, Alpha, Betti and the Megaparsec Universe: On the Topology of the Cosmic Web, in *Lecture Notes in Computer Science*, Vol. 6970 (Springer, 2011) pp. 60–101.

[6] A. G. Sánchez, C. Scóccola, A. Ross, W. Percival, M. Manera, F. Montesano, X. Mazzalay, A. Cuesta, D. Eisenstein, E. Kazin, *et al.*, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological implications of the large-scale two-point correlation function, Monthly Notices of the Royal Astronomical Society **425**, 415 (2012).

[7] E. Papastergis, A. M. Martin, R. Giovanelli, and M. P. Haynes, The Velocity Width Function of Galaxies from the 40% ALFALFA Survey: Shedding Light on the Cold Dark Matter Overabundance Problem, ApJ **739**, 38 (2011), arXiv:1106.0710 [astro-ph.CO].

[8] R. Kennedy, C. Frenk, S. Cole, and A. Benson, Constraining the warm dark matter particle mass with Milky Way satellites, MNRAS **442**, 2487 (2014), arXiv:1310.7739 [astro-ph.CO].

[9] A. V. Tikhonov, S. Gottlöber, G. Yepes, and Y. Hoffman, The sizes of minivoids in the local Universe: an argument in favour of a warm dark matter model?, Monthly Notices of the Royal Astronomical Society **399**, 1611 (2009).

[10] M. R. Lovell, M. Cautun, C. S. Frenk, W. A. Hellwing, and O. Newton, The spatial distribution of Milky Way satellites, gaps in streams, and the nature of dark matter, MNRAS **507**, 4826 (2021), arXiv:2104.03322 [astro-ph.GA].

[11] A. L. Watts, P. J. Elahi, G. F. Lewis, and C. Power, Large-scale structure topology in non-standard cosmologies: impact of dark sector physics, MNRAS **468**, 59 (2017), arXiv:1702.03066 [astro-ph.CO].

[12] Planck Collaboration, Planck 2013 results. XVI. Cosmological parameters, A&A **571**, A16 (2014), arXiv:1303.5076 [astro-ph.CO].

[13] D. J. Eisenstein, I. Zehavi, D. W. Hogg, R. Scoccimarro, M. R. Blanton, R. C. Nichol, R. Scranton, H.-J. Seo, M. Tegmark, Z. Zheng, S. F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F. J. Castander, A. Connolly, I. Csabai, M. Doi, M. Fukugita, J. A. Frieman, K. Glazebrook, J. E. Gunn, J. S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G. R. Knapp, H. Lin, Y.-S. Loh, R. H. Lupton, B. Margon, T. A. McKay, A. Meiksin, J. A. Munn, A. Pope, M. W. Richmond, D. Schlegel, D. P. Schneider, K. Shimasaku, C. Stoughton, M. A. Strauss, M. SubbaRao, A. S. Szalay, I. Szapudi, D. L. Tucker, B. Yanny, and D. G. York, Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies, ApJ **633**, 560 (2005), arXiv:astro-ph/0501171 [astro-ph].

[14] LUX Collaboration, Results from a Search for Dark Matter in the Complete LUX Exposure, Phys. Rev. Lett. **118**, 021303 (2017), arXiv:1608.07648 [astro-ph.CO].

[15] Xenon Collaboration, Dark Matter Search Results from a One Ton-Year Exposure of XENON1T, Phys. Rev. Lett. **121**, 111302 (2018), arXiv:1805.12562 [astro-ph.CO].

[16] Fermi-LAT Collaboration and DES Collaboration, Searching for Dark Matter Annihilation in Recently Discovered Milky Way Satellites with Fermi-Lat, ApJ **834**, 110 (2017), arXiv:1611.03184 [astro-ph.HE].

[17] S. Colombi, S. Dodelson, and L. Widrow, Large-scale structure tests of warm dark matter, Astrophysical Journal **458** (1996).

[18] W. A. Hellwing, C. S. Frenk, M. Cautun, S. Bose, J. Helly, A. Jenkins, T. Sawala, and M. Cytowski, The Copernicus Complexio: a high-resolution view of the small-scale Universe, Monthly Notices of the Royal Astronomical Society **457**, 3492 (2016).

[19] S. Bose, W. A. Hellwing, C. S. Frenk, A. Jenkins, M. R. Lovell, J. C. Helly, and B. Li, The Copernicus Complexio: statistical properties of warm dark matter haloes, Monthly Notices of the Royal Astronomical Society **455**, 318 (2016).

[20] T. Sousbie, The persistent cosmic web and its filamentary structure - I. Theory and implementation, Monthly Notices of the Royal Astronomical Society **414**, 350 (2011).

[21] T. Sousbie, C. Pichon, and H. Kawahara, The persistent cosmic web and its filamentary structure – II. Illustrations, Monthly Notices of the Royal Astronomical Society **414**, 384 (2011).

[22] J. Cisewski, R. A. Croft, P. E. Freeman, C. R. Genovese, N. Khandai, M. Ozbek, and L. Wasserman, Nonparametric 3D map of the intergalactic medium using

the Lyman-alpha forest, Monthly Notices of the Royal Astronomical Society **440**, 2599 (2014).
[23] P. Pranav, H. Edelsbrunner, R. Van de Weygaert, G. Vegter, M. Kerber, B. J. Jones, and M. Wintraecken, The topology of the cosmic web in terms of persistent Betti numbers, Monthly Notices of the Royal Astronomical Society **465**, 4281 (2017).
[24] S. B. Green, A. Mintz, X. Xu, and J. Cisewski-Kehe, Topology of our cosmology with persistent homology, CHANCE **32**, 6 (2019).
[25] P. Pranav, R. Van de Weygaert, G. Vegter, B. J. Jones, R. J. Adler, J. Feldbrugge, C. Park, T. Buchert, and M. Kerber, Topology and geometry of Gaussian random fields I: on Betti numbers, Euler characteristic, and Minkowski functionals, Monthly Notices of the Royal Astronomical Society **485**, 4167 (2019).
[26] X. Xu, J. Cisewski-Kehe, S. B. Green, and D. Nagai, Finding cosmic voids and filament loops using topological data analysis, Astronomy and Computing **27**, 34 (2019).
[27] A. Cole, M. Biagetti, and G. Shiu, Topological Echoes of Primordial Physics in the Universe at Large Scales, arXiv preprint arXiv:2012.03616 (2020).
[28] T. Duong, B. Goud, and K. Schauer, Closed-form density-based framework for automatic detection of cellular morphology changes, Proceedings of the National Academy of Sciences **109**, 8382 (2012).
[29] P. Bendich, J. Marron, E. Miller, A. Pieloch, and S. Skwerer, Persistent homology analysis of brain artery trees, The Annals of Applied Statistics **10**, 198 (2016).
[30] P. Lawson, A. B. Sholl, J. Q. Brown, B. T. Fasy, and C. Wenk, Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology, Scientific reports **9**, 1 (2019).
[31] E. Berry, Y.-C. Chen, J. Cisewski-Kehe, and B. T. Fasy, Functional summaries of persistence diagrams, Journal of Applied and Computational Topology **4**, 211 (2020).
[32] A. Robinson and K. Turner, Hypothesis testing for topological data analysis, Journal of Applied and Computational Topology **1**, 241 (2017).
[33] P. Bubenik, Statistical topological data analysis using persistence landscapes, Journal of Machine Learning Research **16**, 77 (2015).
[34] C. A. Biscio and J. Møller, The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications, Journal of Computational and Graphical Statistics **28**, 671 (2019).
[35] J. Krebs and C. Hirsch, Functional central limit theorems for persistent Betti numbers on cylindrical networks, Scandinavian Journal of Statistics (2021).
[36] V. Springel, S. D. White, G. Tormen, and G. Kauffmann, Populating a cluster of galaxies–I. Results at z= 0, Monthly Notices of the Royal Astronomical Society **328**, 726 (2001).
[37] M. Viel, G. D. Becker, J. S. Bolton, and M. G. Haehnelt, Warm dark matter as a solution to the small scale crisis: New constraints from high redshift Lyman-α forest data, Physical Review D **88**, 043502 (2013).
[38] M. R. Lovell, C. S. Frenk, V. R. Eke, A. Jenkins, L. Gao, and T. Theuns, The properties of warm dark matter haloes, Monthly Notices of the Royal Astronomical Society **439**, 300 (2014).
[39] J. R. Munkres, *Elements of algebraic topology*, Vol. 2 (Addison-Wesley Menlo Park, 1984).
[40] A. Hatcher, *Algebraic topology* (Cambridge University Press, 2002).
[41] H. Edelsbrunner and J. Harer, *Computational topology: an introduction* (American Mathematical Soc., 2010).
[42] A. Zomorodian, Fast construction of the Vietoris-Rips complex, Computers & Graphics **34**, 263 (2010).
[43] P. Niyogi, S. Smale, and S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, Discrete & Computational Geometry **39**, 419 (2008).
[44] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, Stability of persistence diagrams, Discrete & Computational Geometry **37**, 103 (2007).
[45] F. Chazal, V. De Silva, M. Glisse, and S. Oudot, *The Structure and Stability of Persistence Modules* (Springer, 2016).
[46] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, *et al.*, Confidence sets for persistence diagrams, The Annals of Statistics **42**, 2301 (2014).
[47] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer, Fréchet means for distributions of persistence diagrams, Discrete & Computational Geometry **52**, 44 (2014).
[48] E. Munch, K. Turner, P. Bendich, S. Mukherjee, J. Mattingly, J. Harer, *et al.*, Probabilistic Fréchet means for time varying persistence diagrams, Electronic Journal of Statistics **9**, 1173 (2015).
[49] P. Bubenik and P. Dłotko, A persistence landscapes toolbox for topological statistics, Journal of Symbolic Computation **78**, 91 (2017).
[50] P. Bubenik, The persistence landscape and some of its properties, in *Topological Data Analysis* (Springer, 2020) pp. 97–117.
[51] F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman, Stochastic convergence of persistence landscapes and silhouettes, in *Proceedings of the thirtieth annual symposium on Computational geometry* (ACM, 2014) p. 474.
[52] K. R. Mecke, T. Buchert, and H. Wagner, Robust morphological measures for large-scale structure in the universe, arXiv preprint astro-ph/9312028 (1993).
[53] C. Park, P. Pranav, P. Chingangbam, R. Van De Weygaert, B. Jones, G. Vegter, I. Kim, J. Hidding, and W. A. Hellwing, Betti numbers of gaussian fields, Journal of The Korean Astronomical Society **46**, 125 (2013).
[54] Y. Kimura and K. Imai, Quantification of lss using the persistent homology in the sdss fields, Advances in Space Research **60**, 722 (2017).
[55] S. K. Giri and G. Mellema, Measuring the topology of reionization with Betti numbers, Monthly Notices of the Royal Astronomical Society **505**, 1863 (2021).
[56] G. Wilding, K. Nevenzeel, R. van de Weygaert, G. Vegter, P. Pranav, B. J. Jones, K. Efstathiou, and J. Feldbrugge, Persistent homology of the cosmic web–I. hierarchical topology in ΛCDM cosmologies, Monthly Notices of the Royal Astronomical Society **507**, 2968 (2021).
[57] U. Bauer and F. Pausinger, Persistent Betti numbers of random Cech complexes, arXiv preprint arXiv:1801.08376 (2018).
[58] Y. Hiraoka, T. Shirai, K. D. Trinh, *et al.*, Limit theorems for persistence diagrams, Annals of Applied Probability **28**, 2740 (2018).
[59] J. T. Krebs and W. Polonik, On the asymptotic normality of persistent Betti numbers, arXiv preprint arXiv:1903.03280 (2019).

[60] C. A. Biscio, N. Chenavier, C. Hirsch, A. M. Svane, *et al.*, Testing goodness of fit for point processes via topological data analysis, Electronic Journal of Statistics **14**, 1024 (2020).

[61] S. Shandarin, Percolation theory and the cell/lattice structure of the universe, Soviet Astronomy Letters **9**, 104 (1983).

[62] S. Shandarin and I. B. Zeldovich, Topology of the large-scale structure of the universe, Comments on Astrophysics **10**, 33 (1983).

[63] J. R. Gott III, A. L. Melott, and M. Dickinson, The sponge-like topology of large-scale structure in the universe, The Astrophysical Journal **306**, 341 (1986).

[64] J. R. Gott, I. measuring the topology of large-scale structure in the universe, Publications of the Astronomical Society of the Pacific **100**, 1307 (1988).

[65] A. L. Melott, The topology of large-scale structure in the universe, Physics Reports **193**, 1 (1990).

[66] Y.-C. Chen, D. Wang, A. Rinaldo, and L. Wasserman, Statistical analysis of persistence intensity functions, arXiv preprint arXiv:1510.02502 (2015).

[67] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, Persistence images: A stable vector representation of persistent homology, Journal of Machine Learning Research **18** (2017).

[68] D. Yogeshwaran, E. Subag, and R. J. Adler, Random geometric complexes in the thermodynamic regime, Probability Theory and Related Fields **167**, 107 (2017).

[69] Y. Mileyko, S. Mukherjee, and J. Harer, Probability measures on the space of persistence diagrams, Inverse Problems **27**, 124007 (2011).

[70] S. D. Landy and A. S. Szalay, Bias and variance of angular correlation functions, The Astrophysical Journal **412**, 64 (1993).

[71] P. Peebles, The large-scale structure of the universe, Large-Scale Structure of the Universe by Phillip James Edwin Peebles. Princeton University Press (1980).

[72] U. Bauer, Ripser: efficient computation of Vietoris-Rips persistence barcodes, Journal of Applied and Computational Topology 10.1007/s41468-021-00071-5 (2021).

[73] A. J. Benson, C. G. Lacey, C. M. Baugh, S. Cole, and C. S. Frenk, The effects of photoionization on galaxy formation - I. Model and results at z=0, MNRAS **333**, 156 (2002), arXiv:astro-ph/0108217 [astro-ph].

[74] T. Sawala, C. S. Frenk, A. Fattahi, J. F. Navarro, T. Theuns, R. G. Bower, R. A. Crain, M. Furlong, A. Jenkins, M. Schaller, and J. Schaye, The chosen few: the low-mass haloes that host faint galaxies, MNRAS **456**, 85 (2016), arXiv:1406.6362 [astro-ph.CO].

[75] S. Bose, W. A. Hellwing, C. S. Frenk, A. Jenkins, M. R. Lovell, J. C. Helly, B. Li, V. Gonzalez-Perez, and L. Gao, Substructure and galaxy formation in the Copernicus Complexio warm dark matter simulations, MNRAS **464**, 4520 (2017), arXiv:1604.07409 [astro-ph.CO].

[76] A. Baddeley, E. Rubak, and R. Turner, *Spatial point patterns: methodology and applications with R* (CRC press, 2015).

[77] A. J. Baddeley, R. Turner, *et al.*, Spatstat: An R package for analyzing spatial point patterns (2004).

[78] A. Baddeley, R. Turner, E. Rubak, and K. K. Berthelsen, Package 'spatstat', The Comprehensive R Archive Network (2014).

[79] A. Baddeley and R. D. Gill, Kaplan-Meier estimators of distance distributions for spatial point processes, The Annals of Statistics , 263 (1997).