Iterative Probabilistic Hough Transform: Robust Discovery of Filamentary Structures in High-Noise Regimes

Othman Alghamdi
osa210@student.bham.ac.uk
University of Birmingham
School of Computer Science
Birmingham, UK

Marco Canducci
m.canducci@bham.ac.uk
University of Birmingham
School of Computer Science
Birmingham, UK
INAF
Osservatorio Astronomico di Padova
Padova, Italy

Rory Smith
rory.smith@usm.cl
Universidad Técnica Federico Santa María
Departamento de Física
Valparaíso, Chile

Peter Tino
p.tino@bham.ac.uk
University of Birmingham
School of Computer Science
Birmingham, UK

Abstract

Recovering filamentary structures from higher-dimensional point clouds is a central problem in scientific discovery, e.g., in the analysis of large scale dark matter filamentary structure of the Universe. Such structures are often faint, sparsely sampled and submerged in significant levels of noise. Even though observational data generally includes item-wise uncertainty estimates, these are rarely incorporated in a principled manner into manifold-learning methodologies. Finally, the filament detection algorithms usually involve multiple hard-to-interpret parameters whose effect on the results can be difficult to evaluate. To address these issues, we introduce the Iterative Probabilistic Hough Transform (iPHT), a model-based framework for discovering low-signal, spatially coherent noisily sampled filaments embedded in extremely noisy 3D point clouds. The iPHT formulates local structure inference as probabilistic voting over possible line orientations within spatial neighborhoods. We quantify the uncertainty over the existence of a directional signature within each neighbourhood via the entropy of the obtained angular distribution. We then iteratively remove points with highly uncertain neighbourhoods (high entropy), while modifying at each iteration the contribution from each point via a Bayesian update. This provides a principled criterion for suppressing. Importantly, our method is governed by a single physically interpretable parameter allowing explicit control over the inductive bias. We demonstrate on synthetic data with known ground truth that iPHT achieves high-fidelity filament recovery under extreme noise regimes where standard clustering and graph-based methods fail. We then apply iPHT to real world spectroscopic observations of the Large-Scale Structure of the Universe. The method identifies stable, physically meaningful filaments and exhibits greater robustness to parameter variation than the widely-used alternative DisPerSE. We find DisPerSE prone to hallucinating structures under comparable conditions.

CCS Concepts

• **Applied computing** → **Astronomy**; **Mathematics and statistics**; **Physics**; • **Computing methodologies** → **Cluster analysis**; **Dimensionality reduction and manifold learning**.

Keywords

Probabilistic Hough Transform, Filamentary structure detection, Low signal-to-noise, Entropy-based filtering, Curve Fitting, Probabilistic framework, Noise robustness

ACM Reference Format:
Othman Alghamdi, Marco Canducci, Rory Smith, and Peter Tino. 2026. Iterative Probabilistic Hough Transform: Robust Discovery of Filamentary Structures in High-Noise Regimes. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818955

1 Introduction

Clustering is a key tool in unsupervised machine learning for finding patterns in data [13]. However, traditional methods such as K-means and hierarchical clustering struggle with noise, which reduces their accuracy and increases computational cost [12, 30]. To deal with this, several methods have been proposed. CDKNN [15] uses directed k-nearest neighbours to find clusters even when noise is present. DBSCAN [15] looks for dense regions in the data using a neighbourhood radius and a minimum number of nearby points. KdMutual [21] combines density peak and connectivity-based ideas with a filtered single linkage, which helps it handle clusters of different densities without merging them too early.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818955

Torque Clustering (TC) is a physics inspired algorithm that operates without requiring any parameters [34]. It is framed as hierarchical clustering algorithm that builds a full compositional tree of clusters and prunes the tree according to conditions that mimic agglomerative cosmological clustering (based on mass and distance comparisons between leaf clusters).

Despite these advancements, most existing techniques face challenges in accurately determining the number and structure of clusters when noise is high. In this paper, we present a new approach that can identify low-signal structures based on iterative Probabilistic Hough, Model-based Transform (iPHT). PHT offers accuracy and robustness in highly noisy situations, making it especially useful for recognizing one-dimensional structures. The Hough transform [8] was initially designed to detect line elements **(or other parametrized curves)** in pictures. However, it is unreliable when in the presence of noise. In contrast, the PHT extends this framework by leveraging probability distributions and prior knowledge, improving performance in noisy scenarios [36].

Understanding the structure of the Universe at larger scales than the local Group (hosting the Milky Way) is crucial in the comparison of different cosmological models. The inhomogeneities superimposed on a statistically homogeneous background matter distribution are of galaxy (super)clusters, filaments, walls and voids [11]. The collection of these intricate patterns in a global structure at larger scale is referred to as the Cosmic Web [2].

These inhomogeneities grow from small primordial density perturbations in the Λ−CDM model. In overdense regions, self-gravity overcomes the background expansion, leading to the nonlinear collapse of dark matter into haloes and the emergence of the filament–wall–void cosmic web [35].

Cosmological N-body simulations follow the evolution of collisionless dark matter as a large number of particles that sample the underlying matter distribution and interact only through gravity. Gravitational collapse of this particle distribution leads to the formation of dark matter haloes and subhaloes, which are identified with halo-finding algorithms. Galaxy–halo connection models (e.g., halo occupation distributions, abundance matching, and hydrodynamic simulations) associate galaxies with these haloes. Because of this connection, clustering can be used to infer the properties of dark matter and cosmological parameters (see e.g. [31]).

While nodes (hosting clusters and super-clusters of galaxies) in the Cosmic Web define main basins of attraction, filaments (and walls) act as channels directing the flow of the surrounding matter [14]. Thus, tracing cosmic web filaments is essential in understanding the dynamics and distribution of visible and dark matter.

Most filament-finding methods are developed and extensively tested in cosmological simulations [5, 16, 23, 24, 33], with a smaller but growing number of applications to spectroscopic galaxy surveys such as VIPERS [17] and SDSS [18] or specific sub-samples (e.g. around the Fornax-Eridanus complex [20]).

The major filament detection algorithms used within the literature can be categorized into three classes.

Topological Data Analysis: DisPerSe [24] is the most widely used approach in this category and overall in cosmology. While producing skeletons of the filaments directly on point-clouds, its hyperparameter is hard to tune and interpret. Drawing on Topological Data Analysis principles, it is also sensitive to noise corruption [1].

Hessian matrix analysis: Whether applied to the density field (NEXUS [5]) or to the velocity field (COWS [19]), these methods rely on imposing a grid over the volume occupied by the data. Computation of the Hessian is hampered when data is scarce or with systematically varying density, as is the case for observational Cosmic Web. Since both methods need a grid as input rather than raw point clouds, they are not included in the main synthetic benchmark. Instead, we compare them with iPHT separately in Section 3.2, after converting the point cloud to a density grid.

Natural Computation: 1-DREAM [3] (and previous work in multi-manifold learning [4]) exploits emergent behaviour of ant-colonies to highlight over-dense, locally aligned structures. It relies on Principal Component Analysis (PCA) to evaluate the local alignment within local neighbourhoods. However, PCA is sensitive to outliers, thus highly noisy and sparsely populated neighbourhoods have more uncertain orientation than more compact ones. Extreme sparsity reduces the accuracy of 1-DREAM.

We compare the results of iPHT with the ones from DisPerSe, as this is widely adopted as a baseline in the analysis of simulations and observations of the cosmic web. The main contributions of our work are:

- We demonstrate how imposing a model over the directional alignment of points within local neighbourhoods greatly enhances the recovery of noisily sampled manifolds, in the presence of extreme contamination noise.
- We derive a model-fitting technique that relies on the alignment between the model's local tangent space and the distribution of points within the member neighbourhoods.
- We compare iPHT with existing manifold learning and density based clustering techniques, stress-testing their capabilities at extreme noise contamination levels. We find that iPHT recovers much more reliable structures in controlled experiments.
- We show how iPHT can extract detailed, persistent filaments from observations of the Large Scale Structure (LLS) of the Universe and compare to the state-of-the-art in the field (DisPerSe), highlighting the more reliable behaviour of iPHT in the comparison.

2 Methodology Overview

We consider a dataset $\mathcal{D} = \{\mathbf{x}_i \in \mathbb{R}^d \mid i = 1, \dots, N\}$ with $d = 3$, composed of noisy observations of an unknown number of one-dimensional structures (filaments) $\mathcal{S}_j$, each sampled by N_j points and embedded in a noisy background of N_n uniformly distributed points. When $N_n \gg N_j$ filaments are buried within a disruptive amount of contaminating noise (low signal-to-noise regime). Identifying the structures through their over-densities only becomes harder than in higher signal-to-noise situations.

However, filaments are by nature elongated structures, that can be locally approximated by linear segments.

The proposed methodology draws from this property of filaments and recovers them gradually by piecing together (and modelling globally) locally linear "patches" from the noisy data. The

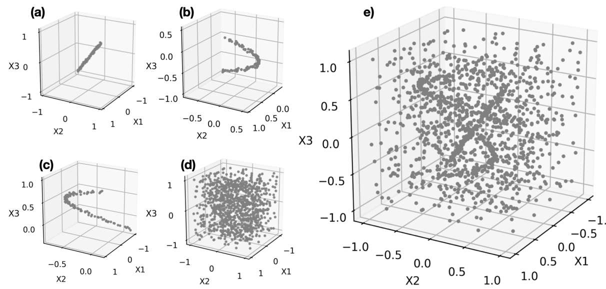

Figure 1: Visualization of filaments in a Noisy Environment.

linearity of an ϵ-neighbourhood is estimated via the Probabilistic, Model-Based Hough Transform (PHT, [36]). Given a point from the point-cloud, we let all points within its R-neighbourhood (from hereon $\epsilon = R$), vote over a set of directions passing through the center, sampling the unit sphere. Votes are obtained by assuming a noise model for the filament thickness and are accumulated to form the Hough accumulator. We thus obtain, for each neighbourhood, a distribution (normalized Hough accumulator) over the angular space (sampled via a fixed set of directions). The entropy of this distribution informs us on the degree to which the neighbourhood presents a one-dimensional alignment. A higher entropy implies broad angular distribution, lacking a clear modality, and thus a clear one-dimensional alignment.

By iteratively filtering out high-entropy points and re-applying PHT, iterative PHT (iPHT) can dig deep into the contamination noise, to enhance the detection of locally aligned neighbourhoods. A degree-two polynomial is fit to surviving points in isolated elongated components with a dedicated procedure that accounts both for the closeness of fit and the alignment between the PHT-detected directions and tangents of the model curve (filament spine). We demonstrate iPHT on a synthetic 3D dataset $\mathcal{D} \in \mathbb{R}^3$ containing three noisily sampled filaments with varying curvatures and alignments, embedded in high uniform noise, as shown in Fig. 1. Each filament is sampled by 100 points and perturbed by uniform noise. The functional forms of the filaments are shown in Appendix B. In addition, the synthetic dataset used in this work, together with the implementation of iPHT, is publicly available on GitHub at https://github.com/OthHadi/iPHT/tree/main/iPHT.

2.1 Iterative Probabilistic Hough Transform (iPHT)

The iPHT involves computing the neighbourhood-wise angular distributions, filtering based on their entropy and iterating until completion. We first describe the theoretical formulation of PHT and then the iterative process.

2.1.1 Model-based PHT - Initialization. A line of length $2R$ passing through a central point in $\mathbf{c} \in \mathbb{R}^3$ can be parameterized using a unit directional vector $\boldsymbol{\omega}$ and a radial coordinate $r \in [-R, R]$: $\mathbf{c} + r \cdot \boldsymbol{\omega}$. The direction space is discretized into N_Ω directions, represented by unit directional vectors $\boldsymbol{\omega}_f$, $f = 1, ..., N_\Omega$, collected in $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_f | f = 1, \ldots, N_\Omega\}$. Each direction $\boldsymbol{\omega}_f$ is radially sampled with N_R points equally spaced in the interval $[-R, R]$. We obtain the set $\mathcal{A} = \{\mathbf{q}_f^a, a = 1, \ldots, N_R; f = 1, \ldots, N_\Omega\}$, of size $|\mathcal{A}| = N_R N_\Omega$.

Given a point $\mathbf{c} \in \mathcal{D}$, its neighborhood is defined as $\mathcal{N}(\mathbf{c}, R) = \mathcal{D} \cap \mathcal{B}(\mathbf{c}, R)$, where $\mathcal{B}(\mathbf{c}, R)$ is the ball of radius R centered at $\mathbf{c}$. Sampled candidate lines passing through the center $\mathbf{c}$ are given by the translated set $\overline{\mathcal{A}} = \{\overline{\mathbf{q}}_f^a = \mathbf{q}_f^a + \mathbf{c}\}_{a,f}$.

Each point $\mathbf{x}_i \in \mathcal{D}$ is viewed as having been generated from a Gaussian distribution centered at a candidate line point $\overline{\mathbf{q}}_f^a$, with spherical covariance $\Sigma = \sigma^2 \mathbf{I}$:

$$p(\mathbf{x}_i | \overline{\mathbf{q}}_f^a, \Sigma) = \mathcal{G}(\mathbf{x}_i | \overline{\mathbf{q}}_f^a, \Sigma). \tag{1}$$

In this study, as parameter σ is fixed for all points it is directly interpreted as the *thickness* of the expected manifolds. Throughout this work $\sigma = R/3$.

To quantify how likely a point $\mathbf{x}_i$ is associated with a particular line, we compute the posterior probability of a candidate point $\overline{\mathbf{q}}_f^a$, given the observation $\mathbf{x}_i$:

$$p(\overline{\mathbf{q}}_f^a | \mathbf{x}_i, \Sigma) = \frac{p(\mathbf{x}_i | \overline{\mathbf{q}}_f^a, \Sigma) p(\overline{\mathbf{q}}_f^a)}{\sum_{k=1}^{N_R} \sum_{l=1}^{N_\Omega} p(\mathbf{x}_i | \overline{\mathbf{q}}_l^k, \Sigma) p(\overline{\mathbf{q}}_l^k)}, \tag{2}$$

which can be interpreted as a probabilistic vote of observation $\mathbf{x}_i$ for direction index f at its a-th grid point.

Assuming the radial and angular priors are independent, the prior over candidate line points $\overline{\mathbf{q}}_f^a$ factorizes as:

$$p(\overline{\mathbf{q}}_f^a) = p(\boldsymbol{\omega}_f) \, p(\overline{\mathbf{q}}^a). \tag{3}$$

The overall vote that point $\mathbf{x}_i$ casts for a direction $\boldsymbol{\omega}_f$ is obtained by marginalizing over the radial grid:

$$p(\boldsymbol{\omega}_f | \mathbf{x}_i, \mathbf{c}). = \sum_{a=1}^{N_R} p(\overline{\mathbf{q}}_f^a | \mathbf{x}_i, \Sigma). \tag{4}$$

Positioning the center at a point $\mathbf{x}_j \in \mathcal{D}$, the directional probabilistic votes from points in its R-neighbourhood, $\mathcal{N}_j \coloneqq \mathcal{B}(\mathbf{x}_j, R) \cap \mathcal{D}$, are mixed in a flat mixture[1], referred to as the Hough accumulator $H(\boldsymbol{\omega}_f; \mathcal{N}_j)$:

$$H(\boldsymbol{\omega}_f; \mathcal{N}_j) = \frac{1}{|\mathcal{N}_j|} \sum_{i=1}^{|\mathcal{N}_j|} p(\boldsymbol{\omega}_f | \mathbf{x}_i, \mathbf{x}_j). \tag{5}$$

2.1.2 Entropy-based noise removal. Note that $H(\boldsymbol{\omega}_f; \mathcal{N}_j)$ is a probability distribution over directions ω_f localized at the point $\mathbf{x}_j$. The normalized entropy $0 \leq E(\boldsymbol{\omega}_f; \mathcal{N}_i) \leq 1$ of this angular distribution,

$$E(H(\cdot; \mathcal{N}_j)) = -\frac{\sum_{f=1}^{N_\Omega} H(\boldsymbol{\omega}_f; \mathcal{N}_j) \log H(\boldsymbol{\omega}_f; \mathcal{N}_j)}{\log(N_\Omega)}, \tag{6}$$

quantifies the degree of (un)certainty about the local directional information in the neighborhood of $\mathbf{x}_j$. Lower entropy values signal possible presence of local directionality, whereas points $\mathbf{x}_j$ with high entropy $E(H(\cdot; \mathcal{N}_j))$ can be discarded as not informative for mining local such structures.

However, in low signal-to-noise regimes, the directional entropy at data points is influenced by the contaminating noise. Thus, the difference in entropy between informative and uninformative neighbourhoods is generally not clear-cut, limiting the possibility of choosing a single threshold above which to remove points.

[1] all points are deemed of equal importance

We adopt instead a conservative iterative approach. After having computed the Hough accumulator in all points of the data set, points with entropy higher than the $(100-\eta)$-th percentile of the entropy values across all points are removed from the data set $\mathcal{D}$. Here η is a parameter regulating the aggressiveness of filtering. Throughout our experiments $\eta = 5$. Thus, only the top 5% of points with highest entropy are removed at each iteration. We then repeat the PHT computation on the residual set. However, the prior over directions is no longer flat. Instead, the prior is now position dependent: at $\mathbf{x}_j$, the prior is set to the Hough accumulator $H(\boldsymbol{\omega}_f; \mathcal{N}_j)$ from the previous step (Bayesian updating).

2.1.3 Iterative prior update. For a given iteration $l > 1$, the estimated Hough accumulator $H(\boldsymbol{\omega}_f; \mathcal{N}_i)^{(l-1)}$ from the previous iteration acts as the prior over the angular space for each neighborhood (over the residual set resulting from iteration $(l-1)$): Eq. (3) is thus reformulated as:

$$p(\overline{\mathbf{q}}_f^a)^{(l)} = H(\boldsymbol{\omega}_f; \mathcal{N}_i)^{(l-1)}\, p(\overline{\mathbf{q}}^a)^{(l)}. \tag{7}$$

We then re-estimate the PHT using eqs. (5) and (6), filter out all points $\mathbf{x}_j$ with entropy $E(H(\cdot; \mathcal{N}_j))$ larger than the re-estimated $95-th$ entropy percentile across all neighbourhoods and update the angular priors via eq. (7).

At each iteration, we check whether filament candidates can be extracted from the residual data. If any are found, they are removed from $\mathcal{D}$ and saved for modelling. The identification and extraction of candidate filaments is described in the next section.

This process is repeated until there are no points left in $\mathcal{D}$. In each iteration, points are removed either because they have high entropy or because they belong to a detected filament. Since we always remove at least $\eta = 5\%$ of points in each iteration, the data set will always be fully depleted and the process will stop.

2.1.4 Directional space discretization. For directional votes to act fairly, the unit directional vectors $\boldsymbol{\omega}_f$ collected in $\boldsymbol{\Omega}$ need to sample the unit sphere uniformly. To that end we employ the regular octahedron technique [10] that guarantees, at each level of refinement, similar area across local linear patches approximating the sphere.

Due to the recursive construction of octahedra (see appendix A), the spatial resolution (quantified through distance between neighboring end-points of $R \cdot \boldsymbol{\omega}_f$ on the sphere of radius R) of the sampled directions is increased at each construction layer. This is non other than the thickness of the detectable filaments. As a value for σ is provided by the user, we recursively sample the spherical surface at radius R. At each level, we evaluate the distance between neighbouring points on the surface produced by the sampling, until we find $\sigma_l \leq \sigma \leq \sigma_u$. Here σ_l is the maximum over all produced distances such that $\sigma_l \leq \sigma$ and σ_u the minimum over the ones where $\sigma_u \geq \sigma$.

The iPHT is run for both resolutions σ_l and σ_u. Section 2.3 describes the methodology used to merge results of the two runs. We find that this merged dual resolution process enhances the filament detection results, when compared with single resolution runs.

2.2 Filament identification and modelling

2.2.1 Detection of filament Candidates. Filament candidates are detected as connected components of the R-neighbourhood graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ over the current dataset $\mathcal{D}$, where each point $\mathbf{x}_i \in \mathcal{D}$ is a vertex and $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{E}$ if $\|\mathbf{x}_i - \mathbf{x}_j\| \leq R$. If a filament is identified as a valid candidate for modelling (see below), its set of points is removed from $\mathcal{D}$ and stored for further spine modelling. At the end of iPHT (depletion of data set $\mathcal{D}$), we end up with a list of promising filament candidates $\mathcal{S}_1, \mathcal{S}_2, \ldots$, with their associated model spines.

Consider now a single connected component $C_\ell = (\mathcal{V}_\ell^c, \mathcal{E}_\ell^c)$ of $\mathcal{H}$. At first, the filament spine parametrized as a second degree polynomial is fitted to the set of its nodes $\mathcal{V}_\ell^c$:

$$\mathbf{f}(t) = \mathbf{W}\mathbf{t}, \quad \mathbf{W} \in \mathbb{R}^{3\times 3}, \quad \mathbf{t} = [1, t, t^2]^\top, \quad t \in [0, 1], \tag{8}$$

with $\mathbf{W}$ as curve coefficients. For each point $\mathbf{x}_i \in \mathcal{V}_\ell^c$, squared distance to the points on the curve is:

$$s(t, \mathbf{x}_i) = \|\mathbf{f}(t) - \mathbf{x}_i\|^2 = (\mathbf{f}(t) - \mathbf{x}_i)^\top (\mathbf{f}(t) - \mathbf{x}_i), \tag{9}$$

which is a 4th degree polynomial in t

$$s(t, \mathbf{x}_i) = \sum_{k=0}^{4} C_k \cdot t^k,$$

with parameters

$$\begin{aligned} C_0 &= \sum_{\kappa=1}^{3} (w_{1,\kappa} - x_i^\kappa)^2, \quad C_1 = \sum_{\kappa=1}^{3} 2w_{2,\kappa}(w_{1,\kappa} - x_i^\kappa), \\ C_2 &= \sum_{\kappa=1}^{3} \left[w_{2,\kappa}^2 + 2w_{3,\kappa}(w_{1,\kappa} - x_i^\kappa) \right], \\ C_3 &= 2\sum_{\kappa=1}^{3} w_{2,\kappa} w_{3,\kappa}, \quad C_4 = \sum_{\kappa=1}^{3} w_{3,\kappa}^2. \end{aligned} \tag{10}$$

Here κ denotes the dimension $\kappa = 1, 2, 3$. The closest point $\mathbf{f}(t_i)$ on the curve to $\mathbf{x}_i$ is found by solving for the minimum of $s(t, \mathbf{x}_i)$:

$$\frac{\partial s(t, \mathbf{x}_i)}{\partial t} = 0, \quad \frac{\partial^2 s(t, \mathbf{x}_i)}{\partial t^2} > 0. \tag{11}$$

The tangent vector to the curve $\mathbf{f}(t)$ at this point is:

$$\dot{\mathbf{f}}(t_i) = \left.\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t}\right|_{t_i} = \mathbf{W}[0, 1, 2t_i]^\top. \tag{12}$$

To validate if a filament candidate is promising, we require data points within $\mathcal{V}_\ell^c$ to show in their Hough accumulator profiles a high alignment with the tangent space of $\mathbf{f}(t)$ evaluated at the closest point to them. The mode $\boldsymbol{\omega}_{f(i)}$ of $H(\boldsymbol{\omega}_f; \mathcal{N}_i)$ for point $\mathbf{x}_i$ is compared to the corresponding tangent vector $\dot{\mathbf{f}}(t_i)$ of $\mathbf{f}(t)$ via

$$\phi\left(\boldsymbol{\omega}_{f(i)}, \dot{\mathbf{f}}(t_i)\right) = \cos^{-1}\left(\frac{\boldsymbol{\omega}_{f(i)}^\top \dot{\mathbf{f}}(t_i)}{\left\|\dot{\mathbf{f}}(t_i)\right\|}\right). \tag{13}$$

For a filament candidate $\mathcal{V}_\ell^c$ to be promising we require that 95% of its points satisfy $\left|\phi\left(\boldsymbol{\omega}_{f(i)}, \dot{\mathbf{f}}(t_i)\right)\right| \leq \chi$. If this is the case the candidate is added to the list $\{\mathcal{S}_j\}_j$ of filament candidates (kept for further modelling) and the set $\mathcal{V}_\ell^c$ is removed from $\mathcal{D}$. We fix the value $\chi = \pi/4$ throughout all our experiments.

To fine-tune the least-squares second degree polynomial to the data in $\mathcal{V}_\ell^c$ we first re-parameterize the curve to lie within the filament's end-points (extremes). The weights of the polynomial curve are then optimized to minimize a custom loss-function that reshapes $\mathbf{f}(t)$ to not only pass close to points from $\mathcal{V}_\ell^c$, but also have its tangent bundle closely aligned with PHT modes of $\mathcal{V}_\ell^c$.

2.2.2 Filament Endpoints Detection. To detect the endpoints $\mathcal{X}_j \in \mathcal{S}_j$ of a candidate filament, we take advantage of the distinctive properties of such structures: their elongation along a dominant direction. We first evaluate principal components of the covariance matrix computed within $\mathcal{S}_j$. All $S = |\mathcal{S}_j|$ points in $\mathcal{S}_j$ are then projected onto the dominant principal component (the one with largest explained variance) $\boldsymbol{\gamma}_1$. The minimum and maximum projected points onto $\boldsymbol{\gamma}_1$ identify the extremes of the filament:

$$\mathbf{m}'_1 = \arg\min_{\mathbf{x}_i \in \mathcal{S}_j} \boldsymbol{\gamma}_1^T \mathbf{x}_i, \quad \mathbf{m}'_2 = \arg\max_{\mathbf{x}_i \in \mathcal{S}_j} \boldsymbol{\gamma}_1^T \mathbf{x}_i. \tag{14}$$

Finally, the selected extreme points in the PCA subspace are mapped back to $\mathbf{m}_1$ and $\mathbf{m}_2$ the corresponding points in the native data space.

2.2.3 Filament modelling. Endpoints $\mathbf{m}_1$ and $\mathbf{m}_2$ of filament $\mathcal{S}_j$ anchor the polynomial curve extremes $\mathbf{f}(t_{\hat{m}})$ and $\mathbf{f}(t_{\hat{n}})$ via a reparameterization $\zeta(\tau)$ that normalizes the parameterization to lie in $[0, 1]$. The two extremes $\mathbf{f}(t_{\hat{m}})$ and $\mathbf{f}(t_{\hat{n}})$ are found following the same optimization strategies in eq. (11). The reparameterization takes the form:

$$\zeta(\tau) = t_{\hat{m}} + (t_{\hat{n}} - t_{\hat{m}})\tau, \tag{15}$$

allowing τ to replace t through an invertible transformation. The parameter t_i that identifies the closest point on the curve from $\mathbf{x}_i$, can then be expressed as $t_i = \zeta(\tau_i)$.

Filament modeling leverages a multi-objective loss function incorporating both geometric fitting and angular alignment:

$$\mathcal{L}(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{f}(\zeta(\tau_i)) - \mathbf{x}_i||^2 - \hat{w}_1 \sum_{i=1}^{N} H\Big(\dot{\mathbf{f}}(\zeta(\tau_i)); \mathcal{N}_i\Big) \tag{16}$$

Here, the first term is the squared error; the second term incorporates local directionality alignment via:

$$\dot{\mathbf{f}}(\zeta(\tau_i)) = \mathbf{W}\ [0, 1, 2\zeta(\tau_i)]^\top (t_{\hat{n}} - t_{\hat{m}}), \tag{17}$$

which yields the tangent vector evaluated at the closest point $\mathbf{f}(\zeta(\tau))$ to $\mathbf{x}_i$. The weight $\hat{w}_1 > 0$ expresses trade-off between the two objectives.

Note that $H(\boldsymbol{\omega}_f; \mathcal{N}_i)$ is only defined for the discrete directions in $\boldsymbol{\Omega}$, so we use the closest one to the tangent $\dot{\mathbf{f}}(\zeta(\tau_i))$:

$$H\Big(\dot{\mathbf{f}}(\zeta(\tau_i)); \mathcal{N}_i\Big) = H\Big(\boldsymbol{\omega}_{f^*}; \mathcal{N}_i\Big), \tag{18}$$

where

$$f^* = \arg\min_{f=1,\dots,N_\Omega} \phi\Big(\boldsymbol{\omega}_f,\ \dot{\mathbf{f}}(\zeta(\tau_i))\Big), \tag{19}$$

and $\phi(\cdot,\cdot)$ is the angular difference between two directions as defined in eq. (13).

To optimize $\mathcal{L}(\mathbf{W})$, simulated annealing is used, updating the weight matrix $\mathbf{W}$ by randomly perturbing one of its elements at each iteration ν during training:

$$o^{(\nu)} = 1 + (-1)^{\varrho}\left[1 - \iota^{\left(1-\frac{\nu}{K}\right)}\right], \tag{20}$$

with $K = 15000$, $\iota \sim \mathcal{U}(0, 1)$. Term $\varrho \sim \mathscr{B}(K, 0.5)$ is a binomially distributed parameter that has values in $\{0, 1\}$, making each iteration a Bernoulli trial. The acceptance probability of transitioning from $\mathbf{W}^{(\nu-1)}$ to $\mathbf{W}^{(\nu)}$ is defined as follows: if $\mathcal{L}(\mathbf{W}^{(\nu)}) \leq \mathcal{L}(\mathbf{W}^{(\nu-1)})$, the new state is accepted with probability 1. Otherwise, it is accepted with probability $\exp\left[\Big(\mathcal{L}(\mathbf{W}^{(\nu-1)}) - \mathcal{L}(\mathbf{W}^{(\nu)})\Big)/J^{(\nu)}\right]$, where $J^{(\nu)}$ is a temperature-like parameter that controls the likelihood of accepting higher-loss states and $J^{(\nu)} = \varpi^{\nu} J^{(0)}$, $\varpi = 0.9995$, and $J^{(0)} = 500$ [9]. This enables global exploration early and fine-tuning later, converging to an optimal spine.

2.3 Model consistency under parameter variation

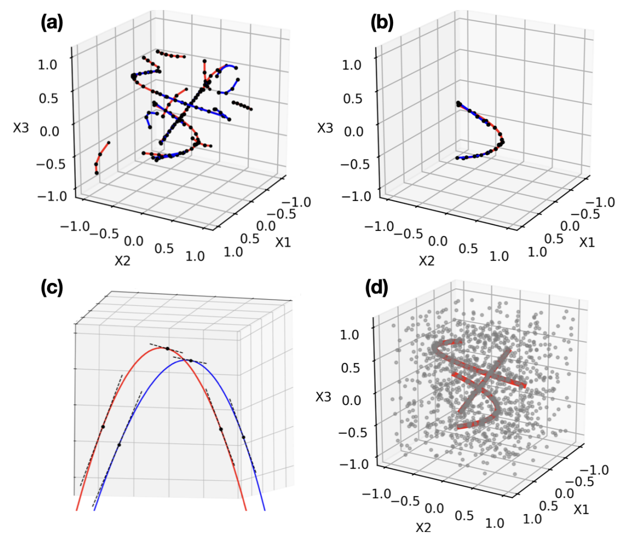

Figure 2: Merging red and blue filaments detected at two thickness scales.

As previously mentioned, discretization of the angular space for the computation of PHT forces the detection and modelling of filaments for two approximate values of filament thickness: σ_l and σ_u. Let us identify by $\{\mathcal{S}_p^l\}_{p=1}^P$ the list of filaments recovered at σ_l and $\{\mathcal{S}_q^u\}_{q=1}^Q$, the one for σ_u (red and blue spines in figures 2a,b,c). Slightly different lists can be recovered in the two cases. However, as this is an effect of the implementation choice, it should not affect the end result. We here propose a method to merge coherent structures and discard discordant ones across the two lists.

Let us denote by $\mathbf{f}_p^l(\zeta(\tau))$, for $p = 1, \dots, P$, and $\mathbf{f}_q^u(\zeta(\tau))$, for $q = 1, \dots, Q$, two spines detected at the thickness values σ_l and σ_u respectively. For both structures, given the parametric equation (8), we can sample a set of points L and U defined as the embedded points corresponding to a set of discrete values for parameter τ: $\mathcal{T}^l = \{\tau_0^l, \dots, \tau_L^l\}$ and $\mathcal{T}^u = \{\tau_0^u, \dots, \tau_U^u\}$ where $\tau_0^l = \tau_0^u = 0$ and $\tau_L^l = \tau_M^u = 1$. Sampling of the parameter τ can be performed such that:

$$\begin{aligned} ||\mathbf{f}_p^l(\zeta(\tau_b^l)) - \mathbf{f}_p^l(\zeta(\tau_{k-1}^l))|| \leq R, \forall b = 1, \dots, L \\ ||\mathbf{f}_q^u(\zeta(\tau_b^u)) - \mathbf{f}_q^u(\zeta(\tau_{k-1}^u))|| \leq R, \forall b = 1, \dots, U. \end{aligned} \tag{21}$$

Tangent vectors are computed using equation (17):

$$\mathbf{a}_p^l(\tau_b^l) = \frac{\mathrm{d}\mathbf{f}_p^l(\zeta(\tau))}{\mathrm{d}\tau}\Big|_{\tau_b^l}, \quad \mathbf{a}_q^u(\tau_b^u) = \frac{\mathrm{d}\mathbf{f}_q^u(\zeta(\tau))}{\mathrm{d}\tau}\Big|_{\tau_b^u}. \tag{22}$$

Spines are merged into $\mathcal{S}_j = \mathcal{S}_j^l \cup \mathcal{S}_j^u$ if at least two consecutive samples meet the two criteria:

$$\text{closeness} : \|\mathbf{f}_p^l(\zeta(\tau_b^l)) - \mathbf{f}_q^u(\zeta(\tau_b^u))\| \leq R \tag{23}$$

$$\text{alignment} : \phi(\mathbf{a}_p^l(\tau_b^l), \mathbf{a}_q^u(\tau_k^q)) \leq \psi, \tag{24}$$

where $\phi(\mathbf{a}_p^l(\tau_b^l), \mathbf{a}_q^u(\tau_k^q))$is the angle between vectors (eq. (13)) and $\psi = \pi/6$. The full process is depicted in fig. 2.

The unified filament $\mathcal{S}_j$ is then reprocessed using PHT with σ_l as the noise model (see 2.1.1) and re-fitted (see 2.2) to derive the final spine. Only filaments meeting closeness and alignment are retained; others are treated as noise.

2.4 Computational Complexity

The main cost is the PHT computation, which runs for every point at each iteration. Building the KDTree costs $O(N \log N)$, and for each point we compute the posterior over its $\bar{N}$ neighbours and $N_R \cdot N_\Omega$ candidate directions, giving a total cost per iteration of:

$$O\left(N \log N + \sum_{j=1}^{N} \bar{N} \cdot N_R \cdot N_\Omega\right). \tag{25}$$

Since $\eta = 5\%$ of points are removed each iteration, N decreases over time and the method always terminates. For sparse data like galaxy surveys $\bar{N} \ll N$, so in practice it runs much faster than the worst case.

3 Experimental Setup

We evaluate the iPHT method on both synthetic and real-world datasets, focusing on its effectiveness in recovering filamentary structures in high-noise environments and comparing its performance with other clustering algorithms.

3.1 Hyperparameter Configuration

The core parameter of iPHT is the neighborhood radius R, which reflects the local scale of the target structures. We set $R = 0.21$. The thickness $\sigma = 0.070$ is approximated via a lower bound $\sigma_l = 0.060$ and upper bound $\sigma_u = 0.075$ to create two angular sampling grids ($N_\Omega = 9, 19, 33$ or 52 total directions). Additional parameters include $N_R = (\lceil R/\sigma_g \rceil \times 2) + 1 = 7$ (radial sampling) and $\eta = 5\%$ (the percentage of points with the highest local entropy that are pruned at each algorithm iteration). iPHT has two kinds of parameters. The first kind is set based on computational considerations. The radial sampling N_R has little effect on results but slows the method down if increased. The simulated annealing parameters ($K = 15000$, $J^{(0)} = 500$, $\varpi = 0.9995$) balance accuracy and speed. The entropy pruning threshold $\eta = 5\%$ controls how many points are removed at each iteration. Using a higher value speeds things up but risks removing useful points too early. The second kind depends on the filament properties, such as the thickness σ, which is taken from the data when the noise level is known, and set to $\sigma = R/3$ otherwise. The alignment threshold $\chi = \pi/4$ accounts for filament curvature and thickness, and we keep it fixed across all experiments in this paper. To remove boundary artifacts (where artificial locally linear patches are created due to the data box), additional points are added in a buffer envelop of the box. These added points are not included in the metric evaluation for any methodology.

Type	Noise	Upper σ	Lower σ	iPHT
Non-Bayesian	10000	$40.0_{30.0}^{53.75}$	$34.0_{28.25}^{43.75}$	$8.0_{\mathbf{7.0}}^{16.5}$
		$0.007_{0.004}^{\mathbf{0.01}}$	$0.006_{0.004}^{0.009}$	$0.002_{0.001}^{0.005}$
Bayesian	10000	$62.5_{53.25}^{73.0}$	$103.5_{51.5}^{118.75}$	$24.0_{\mathbf{20.25}}^{28.5}$
		$0.19_{0.17}^{0.20}$	$0.16_{0.11}^{0.18}$	$0.29_{0.25}^{\mathbf{0.33}}$

Table 1: Ablation study results at noise level 10000. Upper σ and Lower σ are single resolution runs, while iPHT uses the merged dual resolution. Top values are the number of predicted clusters, bottom values are NMI scores in the format $\mathbf{average}_{\mathbf{min}}^{\mathbf{max}}$.

3.2 Synthetic Experiments

We evaluate iPHT and competing methods (KdMutual [21], DBSCAN [15], TC [34], and 1-DREAM [3]) on 30 realizations of the synthetic dataset in Fig. 1, with three filaments (100 points each) embedded in noise levels ranging from 100 to 10,000 points.

Hyperparameters were selected systematically for each method. TC requires no tuning; KdMutual uses prior knowledge for the minimum number of clusters. For both DBSCAN and iPHT, R is varied across 30 values from 0.01 to 0.30. DBSCAN is also evaluated across 19 density thresholds (`minPts` from 2 to 20), while iPHT is tested over 2 combinations of σ_g and entropy threshold η. A consistent elbow pattern in the number of clusters versus R identifies a stable optimal range, to which we restrict the final comparisons (e.g., $R \in [0.02, 0.30]$ for 10,000 noise points).

The comparison is performed with respect to two key metrics: the number of detected clusters (Q^*) and the Normalized Mutual Information (NMI, [25]) between all detected filaments and the ground truth ones. Each methodology is tested 30 times on different realizations of the same ground truth set, 25-th, 50-th and 75-th percentiles over both metrics are reported, for each method, in table 2. For the methodologies requiring hyper-parameter tuning, (iPHT, DBSCAN,1-DREAM) in each run the same range of values over the corresponding hyper-parameters is tested. For these methods, the metrics reported in the table are estimated for the best combination of hyperparameters, identified as the knee-point in the hyperparameters-NMI-Q^* surface. Across all results, the best ones were chosen for each methodology, to be visualized. The recovered structures are shown in Fig. 3.

We study two core components of iPHT. First, the Bayesian prior update uses the Hough accumulator from the previous iteration as the prior for the next iteration. Second, the dual resolution merging combines results from σ_l and σ_u. Table 1 shows what happens when each is removed at a noise level of 10000. Without the Bayesian update, NMI drops from 0.29 to 0.006. The dual-resolution merging also helps by removing filaments that are inconsistent across runs, which reduces false detections.

From the Fig. 2, iPHT consistently predicts cluster counts close to the ground truth and maintains higher, NMI scores across all noise levels. In contrast, DBSCAN, although capable of high performance with specific parameters, often fails with suboptimal ones, leading to low average performance and frequent under-clustering. TC and

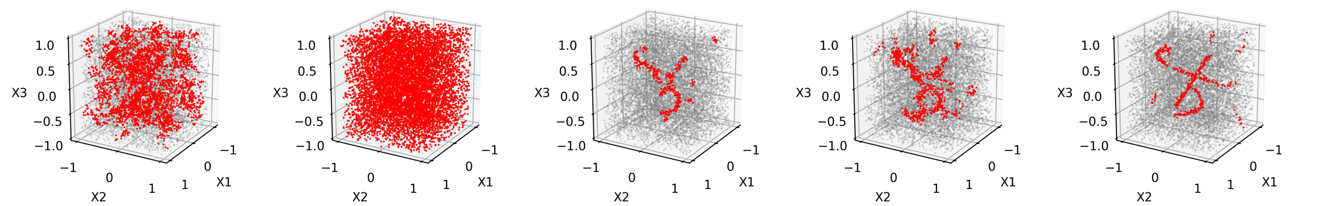

Figure 3: From left to right, results of TC, KdMutual, DBSCAN, 1-DREAM and iPHT in red, over the initial point cloud in grey, when number of noisy points in the volume is 10000.

Noise	TC	KM^*	DBSCAN	$1-D^*$	iPHT
100	4^{4}_{4}	9^{10}_{9}	6^{7}_{5}	4^{4}_{4}	4^{4}_{4}
	$0.88^{0.89}_{0.86}$	$0.84^{0.86}_{0.82}$	$0.87^{0.91}_{0.84}$	$0.77^{0.79}_{0.74}$	$0.90^{\mathbf{0.93}}_{0.87}$
500	7^{11}_{4}	$9^{11}_{7.25}$	7^{8}_{6}	4^{4}_{4}	$5^{5}_{\mathbf{4}}$
	$0.62^{0.67}_{0.55}$	$0.43^{0.45}_{0.41}$	$0.82^{\mathbf{0.84}}_{0.78}$	$0.71^{0.73}_{0.69}$	$0.80^{0.82}_{0.74}$
1000	$17^{22.5}_{13}$	9^{10}_{6}	17^{19}_{13}	4^{4}_{4}	$5.5^{6.75}_{5}$
	$0.39^{0.42}_{0.38}$	$0.27^{0.28}_{0.26}$	$0.71^{0.72}_{0.68}$	$0.65^{0.68}_{0.63}$	$0.71^{\mathbf{0.75}}_{0.67}$
5000	$52^{66.5}_{43.5}$	14^{15}_{12}	$91^{97}_{87.25}$	$11^{13}_{\mathbf{10}}$	15^{18}_{14}
	$0.13^{0.14}_{0.12}$	$0.07^{0.08}_{0.07}$	$0.36^{0.37}_{0.35}$	$0.40^{0.42}_{0.37}$	$0.45^{\mathbf{0.49}}_{0.41}$
10000	$67.5^{81.5}_{60.5}$	$23^{23.75}_{\mathbf{20}}$	$172^{178.75}_{167.25}$	$41.0^{45.5}_{39}$	$24.0^{28.5}_{\mathbf{20.25}}$
	$0.07^{0.08}_{0.07}$	$0.03^{0.04}_{0.03}$	$0.18^{0.19}_{0.17}$	$0.16^{0.17}_{0.16}$	$0.29^{\mathbf{0.33}}_{0.25}$
10000^*	$72.5^{88.8}_{67.0}$	$16.5^{18.0}_{\mathbf{15.0}}$	190^{211}_{167}	$34.5^{37.0}_{32.0}$	$35.0^{37.0}_{30.25}$
	$0.08^{0.09}_{0.08}$	$0.03^{0.03}_{0.02}$	$0.14^{0.15}_{0.13}$	$0.15^{0.16}_{0.13}$	$0.35^{\mathbf{0.37}}_{0.32}$

Table 2: Clustering results on synthetic datasets. Top values in each row are the number of predicted clusters. Bottom values are NMI scores in the format average$^{\mathbf{max}}_{\mathbf{min}}$. KM^* = KdMutual, $1 - D^*$ = 1-DREAM. The row marked 10000^* uses the harder benchmark described in Appendix C.

KdMutual show poor performance under high noise, producing many incorrect clusters.

Fig. 3 shows the best result for each method at 10000 noise points, using the hyperparameter configuration that maximises NMI. TC and KdMutual (the first and second from the left) fail to recover any of the structures. DBSCAN and 1-DREAM (third and fourth) filters remove most of the noise, but DBSCAN breaks filaments into disconnected parts, while 1-DREAM includes additional noise points. iPHT (last plot) correctly identifies most ground-truth filaments while keeping noise contributions low. iPHT also produces some false positives, where noise points accidentally form locally aligned patterns that are mistakenly detected as filaments.

These findings highlight iPHT's robustness and accuracy. Unlike DBSCAN, which is highly sensitive to tuning, iPHT performs reliably across a broad parameter range, making it well-suited for detecting filamentary structures in noisy datasets.

We compare iPHT with DisPerSE, COWS, and NEXUS on the same 30 synthetic datasets, each containing 10,000 noise points. Results are reported as **median**$^{\mathbf{75th}}_{\mathbf{25th}}$ percentile. For DisPerSE, we pick the best F-score across all σ_d values for each realisation, giving $32.6^{37.7}_{29.8}$. For COWS, we first convert the point cloud to a T-web grid following the pipeline in [26], then pick the best parameter combination per realisation, giving $54.6^{57.8}_{52.3}$. iPHT gets $66.1^{69.0}_{65.2}$, which is better than both. For NEXUS, we use NMI instead of F-score, since NEXUS works on a voxel grid and assigns labels to points rather than producing a spine. NEXUS gets $0.0034^{0.0038}_{0.0031}$, while iPHT gets $0.29^{0.33}_{0.25}$ under the same conditions. Overall, COWS and NEXUS both struggle here because they need a density grid to work, which is hard to build reliably when there are so many noise points. iPHT does not have this problem since it works directly on the point cloud.

The row marked 10000^* in Table 2 uses a harder benchmark whose filaments and noise are described in Appendix C. The background noise in this benchmark is non-uniform: it is a mixture of 60% uniformly distributed points and 40% points drawn from a Gaussian distribution centred at the origin $\boldsymbol{\mu} = (0, 0, 0)$ with standard deviation $\sigma_n = 0.5$, which creates a dense noise cluster near the centre of the volume that makes filament detection harder for all methods. In this new benchmark, three filaments are embedded in 10,000 non-uniform background noise points. The first filament follows a cubic S-shaped curve, whose degree-three curvature does not match that of the degree-two spine model. The second filament is a straight line where the points are not evenly distributed along its length, as they are sampled from a Beta distribution. The third filament is a parabolic curve with a thickness that changes along the filament, starting thick at one end and becoming thinner at the other end. Even with these harder conditions, iPHT still achieves a median NMI of 0.35, which is higher than all other compared methods. This result suggests that the good performance of iPHT in the main experiments is not simply because the test data matches the model assumptions, but rather because the probabilistic voting and entropy-based filtering are genuinely robust to this kind of variation.

3.3 The Large Scale Structure of the Universe

We now consider the application of iPHT to the Large Scale Structure (LSS) of the Universe. The data set is obtained by merging different spectroscopic surveys. It was first presented in [29] and used to identify the Large Scale Structure around the Fornax-Eridanus Complex [20]. This catalogue has 78378 galaxies of which, 43480 are brighter than K_s < 11.75 mag (from the 2MASS Redshift Survey), 3627 do not have measured Ks magnitudes, and the remaining from 2M++ are at K_s < 12.5 mag. It is complete for galaxies with

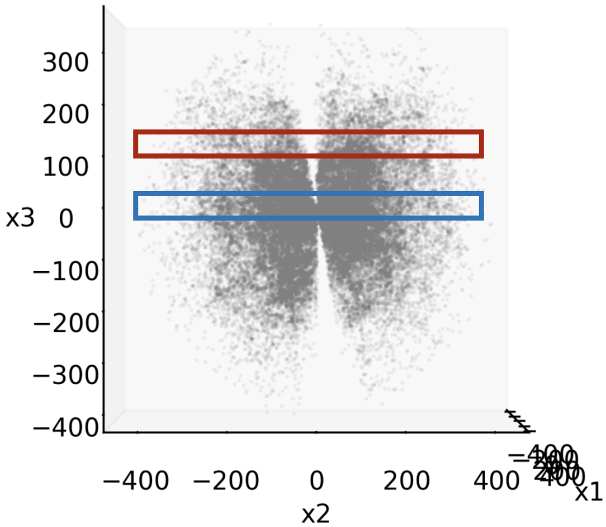

Figure 4: Two slices of thickness 50Mpc each, from The Large Scale Structure of the Universe, for our galaxy sample.

$Ks < 11.5$ mag, which corresponds to $\log_{10}(M_*/M_\odot) = 8.40$ at a distance of 10 Mpc.

The final data set is presented in Fig. 4, along with two slices (red and blue) selected for the comparison of iPHT and Disperse. The two slices are orthogonal to the z-axis, centered at $z = 10$ Mpc and $z = 125$ Mpc and both of thickness 50 Mpc

We test three common hyper-parameters for DisPerSe ($\tilde{\sigma} = 3, 5, 7$) and show all detected filaments at each stage, overlayed in ascending order. The results are shown in figures 5(a) and (d) in shades of blue and with increasing line widths. We follow the recipe explained in [22].

For iPHT, in order to capture the multi-scale properties of the LLS and properly compare with DisPerSe, we run the full methodology for a range of radii, from 2Mpc to 50Mpc, in steps of 2Mpc. To collect results from all radii, we adopt the same technique used in section 2.3. In ascending order, from $r = 2$Mpc to $r = 50$Mpc, we collect all the filaments detected and the associated point-clouds at R_j. These are compared with all filaments at R_{j+1}. Whenever a filament at R_j is included within the ϵ-tube, with $\epsilon = R_{j+1}$, of a filament detected at radius R_{j+1}, the latter is kept and the former discarded. The point-cloud of the filament kept at R_{j+1} will contain the ones from R_j. However, we keep all filaments that do not have intersections across runs, without discarding them (as is done instead in section 2.3).

We obtain a tree of multi-scale structure compositions. The red points in figures 5 (c) and (f) show the union of all points belonging to detected filaments at any radius. The top row shows results for the slice centered at $z = 125$ Mpc, the bottom row, the one at $z = 5$ Mpc. In both slices, the dominant structures at each scale are recovered, while the rest (grey points) are considered noise during the full iPHT process. These points correspond to neighbourhoods with high entropy at any iteration of iPHT and across all radius range.

Fig. 5, panels (b) and (e) show the spines recovered at all scales via the technique proposed in section 2.2.3, as blue lines. The imposition of a degree two polynomial for all filaments (our assumption) might be too strict, as the filaments within the cosmic web show higher curvature structures than predicted.

The results with iPHT are comparable with the ones obtained with DisPerSe (fig. 5 (a) and (d)) in sparse regions. However DisPerSe is highly sensitive to the choice of $\tilde{\sigma}$. While at $\tilde{\sigma} = 3$ (dark blue lines) it tends to align more to existing locally aligned overdensities, at higher values it loses this connection and tends to "hallucinate" structures. For instance, $\tilde{\sigma} = 7$ (light blue curves) shows arbitrarily detected filaments, rarely supported by existing points in the data set. At $\tilde{\sigma} = 3$, it closes the observational gap between the two cones at $z = 0$.

The filaments detected via iPHT seem more anchored on existing points, especially when concentrating towards the inner region of the data set. There, the density of galaxies is higher, being this region closer to the observer and less limited in magnitude (and thus detectability) than the outskirts. In these conditions, where iPHT is designed to provide robust filaments, DisPerSe displays many ramifications and highly non-linear structures, showing the limitation of its topological analysis assumptions.

Furthermore, there is no fixed way or recipe to set hyper-parameter $\tilde{\sigma}$ for DisPerSe a priori, and its lack of interpretability implies that the reliability of the results is subject to the interpretation of the user. Instead iPHT, relying on one dominant physically motivated parameter lets the structure exploration be more robust and objective.

To quantify the visual comparison, we define the *insignificant fraction* as the length-weighted fraction of skeleton segments with no statistically significant galaxy excess, and evaluate it over a grid of two radius multipliers (Appendix D). Figure 6 shows the difference $\xi_1 - \xi_2$ between DisPerSE ($\tilde{\sigma} = 7$) and iPHT. The map is positive across almost the entire $(\tilde{r}, \tilde{b})$ plane, confirming that iPHT places a smaller fraction of its skeleton on regions unsupported by galaxies, by up to ≈ 0.18. This holds across all scanned multiplier combinations and is not an artifact of a particular radius choice.

4 Limitations and Ethical Considerations

We do not envision Ethical Concerns in our work but we provide a brief analysis of the limitations. The computation of iPHT is inherently slower than other methods, due to the computation, for each neighbourhood at a given radius of PHT and the gradual removal. Changing the percentage of points to be filtered out at each iteration may improve speed at the cost of reducing the rate of faint detections. The assumed degree-two polynomial model for filaments may in some cases be too strict. However, it is trivial to modify it to arbitrarily high degree polynomials, as the model is explicitly formulated and tested against the data.

5 Conclusion

We present iPHT, a novel method for detecting filamentary structures in very low signal-to-noise scenarios. At high noise levels, where all compared methods struggle, iPHT still recovers signatures of the ground truth structures. In synthetic experiments, iPHT outperforms DBSCAN, TC, KdMutual, NEXUS, and 1-DREAM in

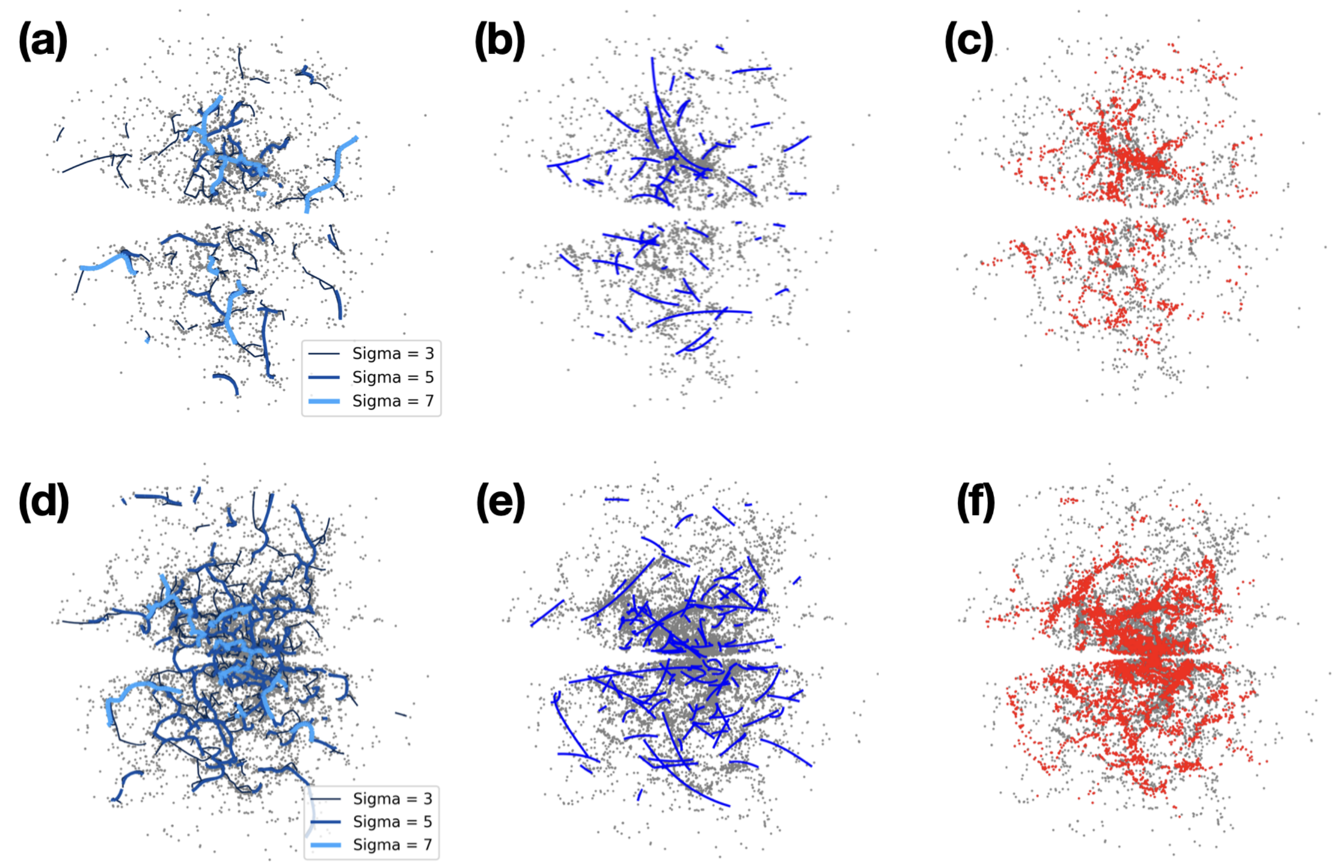

Figure 5: Comparison between iPHT and DisPerSE across 4 selected regions from the SDSS dataset.

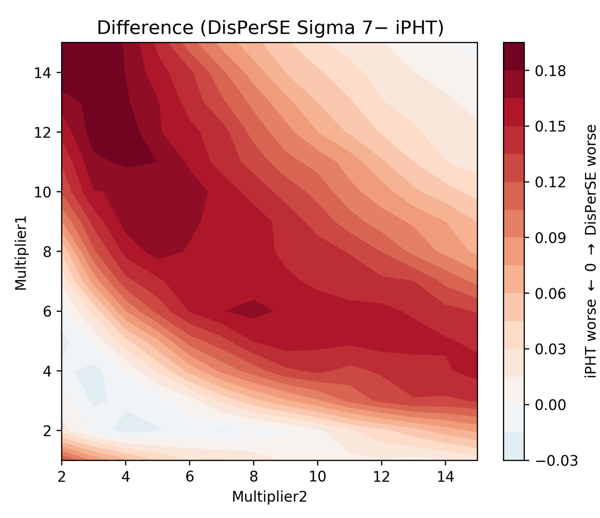

Figure 6: Difference in insignificant fraction (DisPerSE $\tilde{\sigma} = 7$ minus iPHT) as a function of the $\tilde{r}$ (multiplier1) and $\tilde{b}$ (multiplier2).

terms of NMI. Compared to DisPerSE and COWS, iPHT achieves higher accuracy while detecting a much smaller and more reliable set of structures, even when the filaments go beyond the degree-two polynomial assumption of the spine model.

We also applied iPHT to real spectroscopic observations of the Cosmic Web. In sparse regions, the results are comparable to DisPerSE. In higher-density regions, closer to the observer, iPHT recovers more reliable filaments, while DisPerSE tends to produce highly nonlinear structures that are rarely anchored to existing data points. We also show that iPHT is more robust to parameter variation than DisPerSE, whose single hyperparameter is hard to tune and interpret in practice.

Acknowledgments

Othman Alghamdi acknowledges financial support from Taif University [51283/545]. Peter Tino acknowledges financial support from the Engineering and Physical Sciences Research Council [EP/Y031032/1 (EDUCADO)]. Marco Canducci and Peter Tino acknowledge financial support from the Engineering and Physical Sciences Research Council [EP/X025454/1 (ARCANE)]. Rory Smith acknowledges financial support from FONDECYT Regular 2023 [1230441]; and from ANID–MILENIO [NCN2024_112].

References

[1] Petra Awad, Reynier Peletier, Marco Canducci, Rory Smith, Abolfazl Taghribi, Mohammad Mohammadi, Jihye Shin, Peter Tino, and Kerstin Bunte. 2023. Swarm-intelligence-based extraction and manifold crawling along the Large-Scale Structure. *Monthly Notices of the Royal Astronomical Society* 520, 3 (02 2023), 4517–4539. arXiv:https://academic.oup.com/mnras/article-pdf/520/3/4517/49287036/stad428.pdf doi:10.1093/mnras/stad428

[2] Raúl Baier-Soto, Yara Jaffé, Alexis Finoguenov, Christopher P. Haines, Paola Merluzzi, Hugo Méndez-Hernández, Antonela Monachesi, Ulrike Kuchner, Rory Smith, Nicolas Tejos, Cristóbal Sifón, Maria Argudo-Fernández, C. R. Bom, Johan Comparat, Ricardo Demarco, Rodrigo F. Haack, Ivan Lacerna, E. V. R. Lima, Ciria Lima-Dias, Elismar Lösch, C. Mendes de Oliveira, Diego Pallero, Laerte Sodré, Jr., Gabriel S. M. Teixeira, O. Alghamdi, F. Almeida-Fernandes, Stefania Barsanti, Lawrence E. Bilton, M. Canducci, Maiara Carvalho, Giuseppe D'Ago, Alexander Fritz, Fábio R. Herpich, E. Ibar, Hyowon Kim, Sebastian Lopez, Alessia Moretti, L. M. I. Nakazono, D. E. Olave-Rojas, G. B. Oliveira Schwarz, Franco Piraino-Cerda, Emanuela Pompei, U. Rescigno, Boudewijn F. Roukema, V. M. Sampaio, P. Tino, and P. Vásquez-Bustos. 2025. The role of supercluster filaments in shaping galaxy clusters. *Astronomy & Astrophysics* 704, Article A228 (Dec. 2025), A228 pages. arXiv:2510.23311 [astro-ph.CO] doi:10.1051/0004-6361/202556957

[3] Marco Canducci, Petra Awad, Abolfazl Taghribi, Mohammad Mohammadi, Michele Mastropietro, Sven De Rijcke, Reynier Peletier, Rory Smith, Kerstin Bunte, and Peter Tino. 2022. 1-DREAM: 1D Recovery, Extraction and Analysis of Manifolds in noisy environments. *Astronomy and Computing* 41 (2022), 100658. doi:10.1016/j.ascom.2022.100658

[4] Marco Canducci, Peter Tino, and Michele Mastropietro. 2022. Probabilistic modelling of general noisy multi-manifold data sets. *Artificial Intelligence* 302 (2022), 103579. doi:10.1016/j.artint.2021.103579

[5] Marius Cautun, Rien van de Weygaert, and Bernard J. T. Jones. 2012. NEXUS: tracing the cosmic web connection. *Monthly Notices of the Royal Astronomical Society* 429, 2 (12 2012), 1286–1308. arXiv:https://academic.oup.com/mnras/article-pdf/429/2/1286/18456811/sts416.pdf doi:10.1093/mnras/sts416

[6] Dmitry Fuchs. 2016. Geodesics on regular polyhedra with endpoints at the vertices. *Arnold Mathematical Journal* 2 (2016), 201–211.

[7] Álvaro González. 2010. Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. *Mathematical geosciences* 42 (2010), 49–64.

[8] Peter E Hart. 2009. How the Hough transform was invented [DSP History]. *IEEE Signal Processing Magazine* 26, 6 (2009), 18–22.

[9] Darrall Henderson, Sheldon H Jacobson, and Alan W Johnson. 2003. The theory and practice of simulated annealing. *Handbook of metaheuristics* (2003), 287–319.

[10] Adrian Holhoş and Daniela Roşca. 2014. An octahedral equal area partition of the sphere and near optimal configurations of points. *Computers & Mathematics with Applications* 67, 5 (2014), 1092–1107.

[11] Mariana Jaber, Marius Peper, Wojciech A Hellwing, Miguel A Aragón-Calvo, and Octavio Valenzuela. 2023. Hierarchical structure of the cosmic web and galaxy properties. *Monthly Notices of the Royal Astronomical Society* 527, 2 (11 2023), 4087–4099. arXiv:https://academic.oup.com/mnras/article-pdf/527/2/4087/53789031/stad3347.pdf doi:10.1093/mnras/stad3347

[12] Amaninder Kaur, Pankaj Kumar, and Paritosh Kumar. 2010. Effect of noise on the performance of clustering techniques. In *2010 International Conference on Networking and Information Technology*. IEEE, 504–506.

[13] Harpreet Kaur, Swati Malik, Vidhu Baggan, and Shilpi Harnal. 2024. Enhanced K-means Clustering of Tasks and Virtual Machines for Load Balancing in Fog Environment. In *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)*. IEEE, 565–570.

[14] Roi Kugel and Rien van de Weygaert. 2026. Cosmic Web Dynamics: Forces and Strains. *Monthly Notices of the Royal Astronomical Society* (01 2026), stag193. arXiv:https://academic.oup.com/mnras/advance-article-pdf/doi/10.1093/mnras/stag193/66612468/stag193.pdf doi:10.1093/mnras/stag193

[15] Lin-Tao Li, Zhong-Yang Xiong, Qi-Zhu Dai, Yong-Fang Zha, Yu-Fang Zhang, and Jing-Pei Dan. 2020. A novel graph-based clustering method using noise cutting. *Information Systems* 91 (2020), 101504.

[16] Noam I. Libeskind, Rien van de Weygaert, Marius Cautun, Bridget Falck, Elmo Tempel, Tom Abel, Mehmet Alpaslan, Miguel A. Aragón-Calvo, Jaime E. Forero-Romero, Roberto Gonzalez, Stefan Gottlöber, Oliver Hahn, Wojciech A. Hellwing, Yehuda Hoffman, Bernard J. T. Jones, Francisco Kitaura, Alexander Knebe, Serena Manti, Mark Neyrinck, Sebastián E. Nuza, Nelson Padilla, Erwin Platen, Nesar Ramachandra, Aaron Robotham, Enn Saar, Sergei Shandarin, Matthias Steinmetz, Radu S. Stoica, Thierry Sousbie, and Gustavo Yepes. 2017. Tracing the cosmic web. *Monthly Notices of the Royal Astronomical Society* 473, 1 (08 2017), 1195–1217. arXiv:https://academic.oup.com/mnras/article-pdf/473/1/1195/21407912/stx1976.pdf doi:10.1093/mnras/stx1976

[17] N. Malavasi, S. Arnouts, D. Vibert, S. de la Torre, T. Moutard, C. Pichon, I. Davidzon, K. Kraljic, M. Bolzonella, L. Guzzo, B. Garilli, M. Scodeggio, B. R. Granett, U. Abbas, C. Adami, D. Bottini, A. Cappi, O. Cucciati, P. Franzetti, A. Fritz, A. Iovino, J. Krywult, V. Le Brun, O. Le Fèvre, D. Maccagni, K. Małek, F. Marulli, M. Polletta, A. Pollo, L. Tasca, R. Tojeiro, D. Vergani, A. Zanichelli, J. Bel, E. Branchini, J. Coupon, G. De Lucia, Y. Dubois, A. Hawken, O. Ilbert, C. Laigle, L. Moscardini, T. Sousbie, M. Treyer, and G. Zamorani. 2016. The VIMOS Public Extragalactic Redshift Survey (VIPERS): galaxy segregation inside filaments at $z \simeq 0.7$. *Monthly Notices of the Royal Astronomical Society* 465, 4 (11 2016), 3817–3822. arXiv:https://academic.oup.com/mnras/article-pdf/465/4/3817/10254254/stw2864.pdf doi:10.1093/mnras/stw2864

[18] Callum J O'Kane, Ulrike Kuchner, Meghan E Gray, and Alfonso Aragón-Salamanca. 2024. The effect of cosmic web filaments on galaxy evolution. *Monthly Notices of the Royal Astronomical Society* 534, 3 (09 2024), 1682–1699. arXiv:https://academic.oup.com/mnras/article-pdf/534/3/1682/59594327/stae2142.pdf doi:10.1093/mnras/stae2142

[19] Simon Pfeifer, Noam I Libeskind, Yehuda Hoffman, Wojciech A Hellwing, Maciej Bilicki, and Krishna Naidoo. 2022. COWS: a filament finder for Hessian cosmic web identifiers. *Monthly Notices of the Royal Astronomical Society* 514, 1 (May 2022), 470–479. doi:10.1093/mnras/stac1382

[20] Maria Angela Raj, Petra Awad, Reynier F Peletier, Rory Smith, Ulrike Kuchner, Rien van de Weygaert, Noam I Libeskind, Marco Canducci, Peter Tino, and Kerstin Bunte. 2024. Large-scale structure around the Fornax-Eridanus complex. *Astronomy & Astrophysics* 690, Article A92 (Oct. 2024), A92 pages. arXiv:2407.03225 [astro-ph.GA] doi:10.1051/0004-6361/202450815

[21] Frédéric Ros, Serge Guillaume, Mohamed El Hajji, and Rabia Riad. 2020. KdMutual: A novel clustering algorithm combining mutual neighboring and hierarchical approaches using a new selection criterion. *Knowledge-Based Systems* 204 (2020), 106220.

[22] Rory Smith, Ho Seong Hwang, Katarina Kraljic, Paula Calderón-Castillo, Thomas M Jackson, Anna Pasquali, Jihye Shin, Jongwan Ko, Jaewon Yoo, Hyowon Kim, et al. 2023. BCG alignment with the locations of cluster members and the large-scale structure out to 10 R 200. *Monthly Notices of the Royal Astronomical Society* 525, 3 (2023), 4685–4699.

[23] Thierry Sousbie. 2011. The persistent cosmic web and its filamentary structure - I. Theory and implementation. *Monthly Notices of the Royal Astronomical Society* 414, 1 (June 2011), 350–383. arXiv:1009.4015 [astro-ph.CO] doi:10.1111/j.1365-2966.2011.18394.x

[24] Thierry Sousbie, Christophe Pichon, and Hajime Kawahara. 2011. The persistent cosmic web and its filamentary structure – II. Illustrations. *Monthly Notices of the Royal Astronomical Society* 414, 1 (06 2011), 384–403. arXiv:https://academic.oup.com/mnras/article-pdf/414/1/384/3811918/mnras0414-0384.pdf doi:10.1111/j.1365-2966.2011.18395.x

[25] Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3, null (March 2003), 583–617. doi:10.1162/153244303321897735

[26] James Sunseri, Zack Li, and Jia Liu. 2023. Effects of baryonic feedback on the cosmic web. *Physical Review D* 107, 2 (2023), 023514.

[27] Alexander S Szalay, Jim Gray, George Fekete, Peter Z Kunszt, Peter Kukol, and Ani Thakar. 2007. Indexing the sphere with the hierarchical triangular mesh. *arXiv preprint cs/0701164* (2007).

[28] Abolfazl Taghribi, Marco Canducci, Michele Mastropietro, Sven De Rijcke, Kerstin Bunte, and Peter Tino. 2022. ASAP - A sub-sampling approach for preserving topological structures modeled with geodesic topographic mapping. *Neurocomputing* 470 (2022), 376–388. doi:10.1016/j.neucom.2021.05.108

[29] E. Tempel, R. Kipper, A. Tamm, M. Gramann, M. Einasto, T. Sepp, and T. Tuvikene. 2016. Friends-of-friends galaxy group finder with membership refinement. Application to the local Universe. *Astronomy & Astrophysics* 588, Article A14 (April 2016), A14 pages. arXiv:1601.01117 [astro-ph.CO] doi:10.1051/0004-6361/201527755

[30] Michael C Thrun. 2021. Distance-based clustering challenges for unbiased benchmarking studies. *Scientific reports* 11, 1 (2021), 18988.

[31] Risa H. Wechsler and Jeremy L. Tinker. 2018. The Connection Between Galaxies and Their Dark Matter Halos. *Annual Review of Astronomy and Astrophysics* 56, 1 (Sept. 2018), 435–487. doi:10.1146/annurev-astro-081817-051756

[32] Yingxin Wu and Masahiro Takatsuka. 2006. Spherical self-organizing map using efficient indexed geodesic data structure. *Neural Networks* 19, 6-7 (2006), 900–910.

[33] Ziyong Wu, Yu Luo, Wei Wang, Xi Kang, and Renyue Cen. 2024. Cosmological imprints in the filament with DisPerSE. *Monthly Notices of the Royal Astronomical Society* 538, 2 (12 2024), 830–843. arXiv:https://academic.oup.com/mnras/article-pdf/538/2/830/60964817/stae2476.pdf doi:10.1093/mnras/stae2476

[34] Jie Yang and Chin-Teng Lin. 2025. Autonomous clustering by fast find of mass and distance peaks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[35] DH Zhao, HJ Mo, YP Jing, and Gerhard Boerner. 2003. The growth and structure of dark matter haloes. *Monthly Notices of the Royal Astronomical Society* 339, 1 (02 2003), 12–24. arXiv:https://academic.oup.com/mnras/article-pdf/339/1/12/3551163/339-1-12.pdf doi:10.1046/j.1365-8711.2003.06135.x

[36] Hongya Zhao, Hong Yan, et al. 2010. Searching for coexpressed genes in three-color cDNA microarray data using a probabilistic model-based hough transform. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8, 4 (2010), 1093–1107.

A Angular space discretization, implementation details

While the PHT is inherently a continuous probability over all possible directions, a discretization of the angular (and radial) space is needed for its computation. This poses the problem on how to sample points on a spherical surface as regularly as possible, in order to avoid introducing implicit biases in the PHT computation. An initial guess could be to sample regularly the angular space with a finite number of points sampling α and β. For instance, defining the number of points to sample from $\alpha \in [0, \pi[$ as N_α and from $\beta \in [0, \pi[$ as N_β, we would create a regular grid of $N_\Omega = N_\alpha N_\beta$ points. However, these points would sample the spherical surface around the poles more densely than at higher latitudes (see fig. 7 (a)). When computing PHT on this grid, we would be more sensible to neighbourhoods more aligned along the polar axis, than to any direction towards the equator. An alternative could be found in nature, looking at regularities in the growth of floral features. The Fibonacci lattice [7] aims at replicating the patterns found in Nature by defining a sequence of points whose deviation from regularity, across the spherical surface, is minimal. While this approach tends to be a good approximation to a regular grid over the spherical surface, it is non-trivial to adapt it to our case. Indeed, given PHT's purpose in a neighbourhood is to estimate its local alignment, we require symmetry w.r.t. the origin (centre of the neighbourhood). However, when cutting the Fibonacci sequence to one of the two polar hemispheres, its reflection across the origin over the complementary hemisphere does not match the continuity with the original directions. If through each original and reflected point a line is drawn from $-R$ to R, the regularity of the sample is disrupted and the spherical surface's sampling once again departs from uniformity. Geodesic polyhedra [6] are well known, recursive, geometrical structures that closely approximate the spherical surface. In particular, the regularoctahedron has appealing symmetric properties that can be leveraged in the angular space discretization. The paper [10] by Adrian Holhoș and Daniela Roșca presents a novel geometric method for partitioning the sphere into regions of equal area, leveraging the properties of the regular octahedron. The core concept involves constructing an area-preserving map that facilitates a transformation from the unit sphere to the regular octahedron, both centered at the origin. This mapping is essential, as it ensures that the resulting divisions of the sphere maintain equal area, representing a substantial improvement over previous techniques that failed to ensure uniform area partitions. Additionally, the authors employ the inverse of this area-preserving map to formulate uniform and refinable grids on the sphere. Beginning with any triangular uniform and refinable grid, their method facilitates the creation of a hierarchical grid structure that upholds equal area partitions. This innovative approach proves advantageous for computational applications that require consistent area measurements across the sphere. The basic definition of a geodesic octahedron is that of a square-based pyramid, reflected on its base. It is formed by eight triangular faces, each of which is equilateral. The six vertices lie on the circumscribed spherical surface and can act as a regular discretization of the two-dimensional angular space. An initial octahedron can be further refined by regularly increasing its faces via the so called subdivision process. While there exist multiple ways to define this process, we rely on the formulation in [32]. Since we do not require a densely sampling of the angular space, but would like retain a small number of directions where to compute PHT (to save computational time and resources), the gentle increase in octahedral vertices proposed in the cited work is preferred to more intensive versions [27]. In this iterative process, each face, is subdivided in a given number of faces at each iteration.

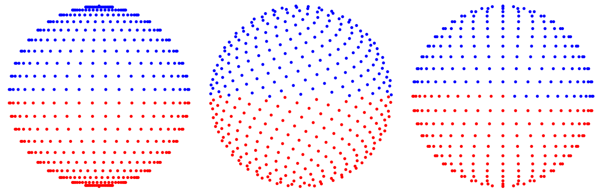

Figure 7: (a) shows a regular grid of $N_\alpha N_\beta$ points. (b) illustrates Fibonacci lattice. (c) shows points mapped from the regular Octahedron to the sphere.

At stage face is divided into a number of triangles according to the stage, v. If we are in the first stage, where $v = 1$, then the number of triangles in one face is equal to v^2 and the number of vertices $\rho = 4v^2 + 2$ [10]. The inverse mapping formulas are used to transfer any grid from the octahedron to the sphere. This method is effective for generating nearly uniform points on the sphere, as demonstrated in Fig. 7 c. We focus on the half of ρ since each vertex and its reflection across the origin represent a single direction within the angular space extending from $-R$ to R. Consequently, the number of directions for each neighbourhood is given by $N_\Omega = \frac{\rho}{2}$. For each point on the sphere, the angles α and β are equal to the elements $\alpha_{(f,1)}$ and $\alpha_{(f,2)}$ of $\overline{\mathbf{a}}_f$, respectively.

The subdivision process is only characterized by the number of iterations and the inverse mapping used to map newly created vertices onto the spherical surface. Stopping the subdivision process at a given iteration implies assuming a certain thickness of the detectable stream. This can be estimated as the minimum geodesic distance (see e.g. [28]) between all neighbouring pairs of points on the spherical surface. We will refer to the stream's thickness as σ_g. The user can provide assumptions about the detectable streams by imposing a value for their thickness. Since the subdivision process can only create a finite (quantized) number of points on a sphere and this number determines the spherical points mutual separation, the estimated mutual separation between spherical points is bound to be different from the provided σ_g. In order to match the user's imposition, we create two sets of points on the sphere by finding the closest lower σ_l and upper σ_u to σ_g, such that $\sigma_l < \sigma_g < \sigma_u$. This is achieved simply by stopping the subdivision process at two consecutive iterations. While this choice allows for more fidelity to the user's requirement, it also poses implementation challenges. In particular, the whole methodology has to be performed for both σ_g, σ_l and σ_u and a technique needs to be implemented for matching the detected structures (see sec. 2.3).

B Filament functional forms

The synthetic data set $\mathcal{D} \in \mathbb{R}^3$ is composed of three noisily sampled streams with different curvatures and alignments across the occupied volume. Stream $\mathcal{S}_1$ is a straight line across the diagonal of the occupied volume. Streams $\mathcal{S}_2$ and $\mathcal{S}_3$ are parabolic streams with different curvatures and orientations. The functional forms are displayed in the sequence shown below:

$$\begin{aligned}
\mathbf{s}_1(g_1) &= \begin{cases} x^1 = g_1 \sim \mathcal{U}([-1,1]) \\ x^2 = g_1 \\ x^3 = g_1 \end{cases} \\
\mathbf{s}_2(g_2,g_3) &= \begin{cases} x^1 = g_2 \sim \mathcal{U}([-0.35,1]) \\ x^2 = -2g_2^2 + 2g_2 \\ x^3 = g_3 \sim \mathcal{U}([-1,0.5]) \end{cases}, \\
\mathbf{s}_3(g_1,g_3) &= \begin{cases} x^1 = g_1 \\ x^2 = g_1^2 - 0.5 \\ x^3 = g_3 \end{cases} \\
\mathcal{S}_j &= \mathbf{s}_j + \mathcal{U}([-0.02,0.02]^3)
\end{aligned} \tag{26}$$

where we define by $[-0.02, 0.02]^3$ the uniform multivariate distribution within the Cartesian product of segments $[-0.02, 0.02] \times [-0.02, 0.02] \times [-0.02, 0.02]$ in directions x^1, x^2 and x^3 respectively.

C Hard filament functional forms

The hard synthetic data set $\mathcal{D} \in \mathbb{R}^3$ is composed of three noisily sampled filaments whose properties go beyond the degree-two polynomial assumption of the spine model. Stream $\mathcal{S}_4$ is a cubic S-shaped curve whose degree-three curvature strictly mismatches the degree-two spine model. Stream $\mathcal{S}_5$ is a straight line where the points are not uniformly distributed along its length. Stream $\mathcal{S}_6$ is a parabolic curve with a thickness that varies continuously along the filament. The functional forms are shown below:

$$\begin{aligned}
\mathbf{s}_4(g_4) &= \begin{cases} x^1 = 0.53\, g_4 - 0.48, \quad g_4 \sim \mathcal{U}([-0.85,\ 0.85]) \\ x^2 = 0.48\,(4g_4^3 - 3g_4) - 0.45 \\ x^3 = 0.48\, g_4 - 0.48 \end{cases} \\
\mathbf{s}_5(g_5) &= \begin{cases} x^1 = g_5 - 0.9, \quad g_5 \sim \mathrm{Beta}(0.4,\ 2.0) \\ x^2 = 0.5\,(g_5 - 0.9) + 0.65 \\ x^3 = -0.6\,(g_5 - 0.9) + 0.35 \end{cases} \\
\mathbf{s}_6(g_6) &= \begin{cases} x^1 = 0.5\, g_6 + 0.4, \quad g_6 = \mathrm{linspace}(-0.3,\ 0.9,\ N) \\ x^2 = 0.45\,(-1.5g_6^2 + 2g_6) + 0.35 \\ x^3 = g_6 \end{cases} \\
\mathcal{S}_j &= \mathbf{s}_j + \boldsymbol{\delta}_j
\end{aligned} \tag{27}$$

For $\mathcal{S}_4$ and $\mathcal{S}_5$, the perturbation is uniform: $\boldsymbol{\delta}_j \sim \mathcal{U}([-0.02,\ 0.02]^3)$. For $\mathcal{S}_6$, the perturbation has a spatially varying magnitude: $\boldsymbol{\delta}_6 = \boldsymbol{u} \odot \boldsymbol{\lambda}$, where $\boldsymbol{u} \sim \mathcal{U}([-1,\ 1]^3)$ and $\boldsymbol{\lambda} = \mathrm{linspace}(0.08,\ 0.005,\ N)$ decreases linearly from one end of the filament to the other, producing a tapered thickness. The non-uniform background noise is a mixture of two components: 60% of points are drawn uniformly from $[-1, 1]^3$, and 40% are drawn from a Gaussian distribution with mean $\boldsymbol{\mu} = (0, 0, 0)$ and standard deviation $\sigma_n = 0.5$, clipped to $[-1, 1]^3$. This creates a dense noise concentration near the centre of the volume, which makes the separation of filament points from noise harder for all methods.

D Galaxy-support measure

Without a ground-truth filament map, we assess each skeleton against the galaxy field directly: a genuine filament segment sits on a transverse overdensity, while a hallucinated segment shows no galaxy excess. We quantify this per segment and summarise each skeleton by the fraction of its total length that carries no significant excess.

Per-segment geometry. Each skeleton is a set of straight segments. For iPHT, each quadratic spine curve is sampled at $\tilde{P} = 15$ unequal length segments, giving $\tilde{P} - 1$ consecutive segments. For DisPerSE ($\tilde{\sigma} = 7$) the segments are taken directly from the `.segs` output file. Segments keep their native, unequal lengths.

A segment has endpoints $\tilde{\mathbf{e}}$ and $\tilde{\mathbf{b}}$, length $\tilde{L} = \|\tilde{\mathbf{b}} - \tilde{\mathbf{e}}\|$, and unit direction $\hat{\mathbf{u}} = (\tilde{\mathbf{b}} - \tilde{\mathbf{e}})/\tilde{L}$. A galaxy at position $\mathbf{p}$ has longitudinal coordinate and perpendicular distance

$$\tilde{\ell} = (\mathbf{p} - \tilde{\mathbf{e}}) \cdot \hat{\mathbf{u}}, \qquad \tilde{d} = \left\|(\mathbf{p} - \tilde{\mathbf{e}}) - \tilde{\ell}\,\hat{\mathbf{u}}\right\|, \tag{28}$$

and is attributed to the segment only if $0 \le \tilde{\ell} \le \tilde{L}$.

Radii. We define an inner cylinder of radius $\tilde{R}$ and an outer annulus of radius $\tilde{Q}$, both scaling with the segment length $\tilde{L}$:

$$\tilde{R} = \tilde{r}\,\tilde{L}, \qquad \tilde{Q} = \tilde{b}\,\tilde{R}, \tag{29}$$

where $\tilde{r} \in \{1, \dots, 15\}$ and $\tilde{b} \in \{2, \dots, 15\}$ are the multipliers scanned on the axes of Fig. 6. By construction $\tilde{Q} > \tilde{R}$ for all $\tilde{b} > 1$. The 3D cylindrical shell volume between two perpendicular radii $\tilde{c}_1$ and $\tilde{c}_2$ around a segment of length $\tilde{L}$ is

$$\tilde{V}(\tilde{c}_1, \tilde{c}_2, \tilde{L}) = \pi\,(\tilde{c}_2^2 - \tilde{c}_1^2)\,\tilde{L}. \tag{30}$$

Per-segment significance. Let $\tilde{N}$ be the galaxy count inside the inner cylinder ($\tilde{d} < \tilde{R},\ 0 \le \tilde{\ell} \le \tilde{L}$) and $\tilde{M}$ the count in the annulus ($\tilde{R} \le \tilde{d} < \tilde{Q},\ 0 \le \tilde{\ell} \le \tilde{L}$). The local under the assumption of Poisson distributed counts is

$$\tilde{n} = \frac{\tilde{M}}{\tilde{V}(\tilde{R}, \tilde{Q}, \tilde{L})}, \tag{31}$$

or the global survey density if the annulus is empty. The per-segment significance is

$$\tilde{S} = \frac{\tilde{N} - \tilde{\lambda}}{\sqrt{\tilde{\lambda}}}, \qquad \tilde{\lambda} = \tilde{n}\,\tilde{V}(0, \tilde{R}, \tilde{L}). \tag{32}$$

Insignificant fraction. A segment is *insignificant* when $\tilde{S} < \tilde{a} = 3$. The length-weighted insignificant fraction $\tilde{\xi}$ is

$$\tilde{\xi} = \frac{\sum_{j:\tilde{S}_j < \tilde{a}} \tilde{L}_j}{\sum_j \tilde{L}_j}, \tag{33}$$

with $\tilde{L}_j$ the length of segment j. We evaluate Eq. (33) on the full grid of $(\tilde{r}, \tilde{b})$ for both methods using the same galaxy catalogue and the same $\tilde{a}$, giving ξ_1 for DisPerSE and ξ_2 for iPHT. The difference $\xi_1 - \xi_2$ is shown in Fig. 6.