



Full length article

1-DREAM: 1D Recovery, Extraction and Analysis of Manifolds in noisy environments[☆]M. Canducci^{a,*}, P. Awad^b, A. Taghribi^c, M. Mohammadi^c, M. Mastropietro^d, S. De Rijcke^d, R. Peletier^b, R. Smith^e, K. Bunte^c, P. Tiño^a^a University of Birmingham, School of Computer Science, B15 1TT, Birmingham, United Kingdom^b University of Groningen, Kapteyn Astronomical Institute, 9747 AD Groningen, The Netherlands^c University of Groningen, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, 9700 AK Groningen, The Netherlands^d Ghent University, Department of Physics and Astronomy, Krijgslaan 281, S9, B-9000 Gent, Belgium^e Universidad Técnica Federico de Santa María, Avenida Vicuña Mackenna 3939, San Joaquín, Santiago, Chile

ARTICLE INFO

Article history:

Received 9 March 2022

Accepted 23 September 2022

Available online 29 September 2022

Keywords:

Methods: N-body simulations

Methods: data analysis

Methods: statistical

Galaxies: Dwarf

(Cosmology:) large-scale structure of universe

(Galaxy:) globular clusters: individual (Omega-Centauri)

ABSTRACT

Filamentary structures (one-dimensional manifolds) are ubiquitous in astronomical data sets. Be it in particle simulations or observations, filaments are always tracers of a perturbation in the equilibrium of the studied system and hold essential information on its history and future evolution. However, the recovery of such structures is often complicated by the presence of a large amount of background and transverse noise in the observation space. While the former is generally considered detrimental to the analysis, the latter can be attributed to measurement errors and it can hold essential information about the structure. To further complicate the scenario, one-dimensional manifolds (filaments) are generally non-linear and their geometry difficult to extract and model. Thus, in order to study hidden manifolds within the dataset, particular care has to be devoted to background noise removal and transverse noise modeling, while still maintaining accuracy in the recovery of their geometrical structure. We propose 1-DREAM: a toolbox composed of five main Machine Learning methodologies whose aim is to facilitate manifold extraction in such cases. Each methodology has been designed to address particular issues when dealing with complicated low-dimensional structures convoluted with noise and it has been extensively tested in previously published works. However, for the first time, in this work all methodologies are presented in detail, joint within a cohesive framework and demonstrated for three particularly interesting astronomical cases: a simulated jellyfish galaxy, a filament extracted from a simulated cosmic web and the stellar stream of Omega-Centauri as observed with the GAIA DR2. Two newly developed visualization techniques are also proposed, that take full advantage of the results obtained with 1-DREAM. This contribution presents the toolbox in all its details and the code is made publicly available to benefit the community. The controlled experiments on a purposefully built data set prove the accuracy of the pipeline in recovering the real underlying structures.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Physical structures with filament-like shapes are found in a wide variety of astrophysical domains. In this work we focus on examples of filamentary (stream-like) structures, such as jellyfish

galaxy tails, filaments of the cosmic web, and the tidal tails of the globular cluster Omega-Centauri (ω -Cen). Since these systems have a high level of complexity, the processes by which they form and evolve are yet to be fully explored and understood, hence the study of filamentary astronomical structures requires powerful computational methods for their detection, modeling and analysis.

Each of these examples shows clear morphological significance to this work in that they all have a stream-like structure, but our interest in these examples is also inspired from their astrophysical importance. The interest in jellyfish galaxies stems from their importance in studying the evolution of galaxies in dense environments and in determining the details of environmental influences on galaxies (Boselli and Gavazzi, 2006; Grossi, 2018). In more details, when a dwarf galaxy enters the environment of

[☆] This project has received financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721463 to the SUNDIAL ITN Network.

* Corresponding author.

E-mail addresses: M.Canducci@bham.ac.uk (M. Canducci), p.awad@rug.nl (P. Awad), abolfazl.taghribi@gmail.com (A. Taghribi), m.mohammadi@rug.nl (M. Mohammadi), michele.mastropietro@ugent.be (M. Mastropietro), sven.derijcke@ugent.be (S. De Rijcke), r.f.peletier@astro.rug.nl (R. Peletier), rorysmith274@gmail.com (R. Smith), kerstin.bunte@googlemail.com (K. Bunte), p.tino@cs.bham.ac.uk (P. Tiño).

a galaxy cluster, its interstellar matter (gas and dust) is often blown out of its body producing a long tail of the constituting matter. Other related studies have been dedicated to examining the effect of ram pressure on the structure and dynamics of these galaxies (Mori and Burkert, 2000; Mayer et al., 2006; Roediger and Brüggén, 2008; Tonnesen and Bryan, 2012; Roediger et al., 2015; Steinhäuser et al., 2016; Yun et al., 2019; Steyrleithner et al., 2020). The second example is that of the cosmic web that is a network of structures that naturally form as a result of gravitational instability within cosmological volumes, and that have provided great insight into gravitational structure formation, cosmological models, as well as the nature of dark matter and dark energy (Park and Lee, 2007; Platen et al., 2008; Lee and Park, 2009; Lavaux and Wandelt, 2010; Bos et al., 2012; Sutter et al., 2015; Pisani et al., 2015). In addition, studies of the cosmic web have given important quantitative measures which improved our understanding of the formation and evolution of galaxies (Hahn et al., 2007a; Hahn, 2009; Cautun et al., 2014). As a final example, we study a stellar stream located in the halo of the Milky Way. Stellar streams are of great importance as they are imprints of past merger events in the Milky Way's formation history (Johnston et al., 1996; Majewski, 1999; McConnachie et al., 2009; Martínez-Delgado et al., 2010; Vera-Casanova et al., 2021). Since the dynamics of these merger events are largely dictated by the galaxy's gravitational potential, tidal debris including stellar streams belonging to globular clusters have also been used as probes of the galactic potential as they move through it (Kupper et al., 2015; Pearson et al., 2015; Thomas et al., 2017, 2018; Bonaca and Hogg, 2018; Malhan and Ibata, 2019).

Extracting topological information from point clouds of discrete data forming the above structures is a difficult task, whether these data sets are provided by N -body simulations or by observational surveys. For example, algorithms studying the cosmic web should keep in mind its anisotropic, hierarchical nature which forms different morphological structures spanning over six orders of magnitude in density (Cautun et al., 2014). Similarly for the other studied objects, structure detection and learning algorithms have to face problems including the very large numbers of high-dimensional data points as well as the handling of noise or outliers that affect the results of manifold learning and dimensionality reduction. That being said, several methods and algorithms have been developed aiming to extract astrophysical information provided by the structures previously mentioned. Starting with jellyfish galaxies, morphometry, or the quantitative study of morphological properties, has been used to study ram pressure stripping in McPartland et al. (2016) and as a probe of the evolution of these galaxies in Roman-Oliveira et al. (2021). A myriad of algorithms has also been developed for the tracing and studying of the various components of the cosmic web (e.g. Multiscale Morphology Filter [MMF], Aragón-Calvo et al. (2007); ORIGAMI, Falck et al. (2012); NEXUS+, Cautun et al. (2013); Minimum Spanning Tree [MST], Alpaslan et al. (2014); Bisous, Tempel et al. (2016)). We refer to Libeskind et al. (2018) for a comparison between many of these algorithms. As for stellar streams, several techniques have been used for their detection and analysis, exemplified by : the Matched Filter (MF; Rockosi et al. (2002), Balbinot et al. (2011)), detecting co-moving groups of stars (Williams et al., 2011; Arifanto and Fuchs, 2006; Dufau et al., 2006), the Streamfinder algorithm (Malhan and Ibata, 2018), and several others. Despite the large number of techniques used to study these systems, the need to keep up with the growing size of astronomical data sets and with the many complexities of astrophysical systems is ever-present. Therefore, the development of new Machine Learning algorithms with astrophysical applications creates great potential towards handling larger amounts of data as is needed for the exploration of the galactic halo and stellar streams belonging to it, as well as towards

handling all the challenging properties of complex structures as required for the analysis of jellyfish tails and the cosmic web.

When studying noisy low-dimensional manifolds, it is often necessary to distinguish between two kinds of noise. In the case of point-clouds, these can be defined as background and transverse noise, however the actual distinction between the two is potentially difficult. While background noise is usually referred to as a contaminant to the data, corrupting information hidden within it, transverse noise may hold useful information about the sub-structures. Due to the local overlap between these two types of noise in the vicinity of the sub-structures, discerning between the two contributions can be a difficult task. Nevertheless, since most Manifold Learning techniques are not designed to deal with such "corruption" of the data, it is generally good practice to address this problem before their application. A number of filtering and denoising techniques have been devised that aim at reducing the noise over a point-cloud (see Han et al., 2017 for an extensive review). However, these seem to work efficiently only in mild cases, where the density of background noise is far lower than the one of the structure. On the other hand, the ant system (Dorigo et al., 1999) and ant-colony system (Gambardella and Dorigo, 1996) have been applied efficiently to denoising in presence of a moderate amount of background noise (Chu et al., 2004). Nevertheless, due to the adopted distance metric (Euclidean) and the computational complexity, these methods can be ineffective on large data sets presenting noisy, non-linear manifolds.

When manifolds are expected within a noisy point-cloud, more direct methodologies are usually applied in order to let their mean curve/surface emerge from the noise. In this class of methodologies, there is generally no distinction between background and transverse noise. Local smoothing has been successfully applied to noisy manifolds with a mild level of noise, using different approaches. In Park et al. (2004) a weighted version of Principal Component Analysis (PCA) is applied to local neighborhoods, estimated in terms of a point's k -nearest neighbors. Then, points in the neighborhood are projected onto the hyperplane defined by the weighted PCA. A different approach is presented in Chen et al. (2006), where a linear error-in-variables (EIV) is estimated for each local patch and the smoothed coordinates of noise-less points are derived. Global coordinates are then recomputed for each point, merging partially overlapping neighborhoods.

A different class of methods aims at projecting neighboring points to the locally estimated tangent space to noisy manifolds. The tangent space estimation can be either performed on the high-dimensional sample (Hao et al., 2017) or the low-dimensional projection (e.g. Yao and Xia, 2019). Manifold Blurring Mean Shift (MBMS, Wang and Carreira-Perpiñán, 2010) adopts this formalism, by gradually moving neighboring points along the orthogonal direction to the manifold. Again, the local tangent space estimation is performed via PCA, in small neighborhoods centered on each point. A slightly different approach is presented in Lyu et al. (2019), where Non-linear Robust PCA is introduced. Here, local patches are decomposed into a low-rank and a sparse component that account for the tangent space and noise information of the patch, respectively.

Another strand of work uses a diffusion formulation over the noisy cloud to enhance the manifold's spine. In particular, Hein and Maier (2007) first construct an asymmetric k -nearest neighborhood graph of the point cloud and derive its graph Laplacian. This serves as a generator for the diffusion process, which is solved in terms of a differential equation. A similar approach but with a physically inspired formulation is presented in Wu et al. (2018). In contrast to the previous methodology, this method does not use a graph to represent the data and gradually moves points towards high density regions in the data.

As previously mentioned, these methodologies are used as a pre-processing step when the data is corrupted by noise. Further steps are necessary to recover explicit formulations of the hidden manifolds and their low-dimensional representations. This branch of work falls within the scope of Manifold Learning and has been addressed in a variety of different methodologies. The stepping stones to this field are *Locally Linear Embedding* (LLE, Roweis and Saul, 2000) and *Complete isometric feature mapping* (Isomap, Tenenbaum et al., 2000). The methodologies have been further refined and new algorithms defined such as Laplacian- and Hessian-eigenmaps (Belkin and Niyogi, 2001; Donoho and Grimes, 2003), Local Tangent Space Alignment (LTSA, Huo et al., 2008) and Riemannian Manifold Learning (Lin and Zha, 2008). However, their performance is often hampered by the presence of noise. Aided by a precise formulation of transverse noise, Generative Topographic Mapping (GTM, Bishop et al., 1998b) solves this issue by modeling the manifold as a Gaussian mixture, having centers constrained to lie on the image of a low-dimensional unit interval (of the same dimension as the manifold) smoothly embedded in the ambient space.

To further complicate the scenario, information may be lying on multiple noisy manifolds in real-world data sets. Generalizations of existing methods to this case were proposed accordingly, giving rise to algorithms such as *Multi Manifold Isomap* (M-Isomap Fan et al., 2016), *Multi-Manifold LLE* (MM-LLE (Hettiarachchi and Peters, 2015) and *Hierarchical-GTM* (Tino and Nabney, 2002). However, carrying the same assumptions and design of their predecessors, they suffer from the same problems (e.g. pre-defined manifold's intrinsic dimension, transverse noise, topologically difficult manifolds). Other techniques gave new perspectives on the problem; examples are *Sparse Manifold Clustering and Embedding* (SMCE Elhamifar and Vidal, 2011) and *Manifold Deflation* (Ting and Jordan, 2020), where manifolds (and their low-dimensional representations) are recovered by means of a graph representation of data. More recently, new mathematical tools have been developed and widely used in multiple fields, although mainly constrained to work in three dimensions (3D point clouds for surface recovery). *Computational Geometry* (Boissonnat et al., 2018) represents manifolds as Simplicial Complexes and refines these representations via triangulations and filtrations. Again though, transverse noise may heavily affect the results and hampers their accuracy.

The general assumption for these methodologies is that the intrinsic dimensionality of the manifolds is known a priori. This is not often the case, so much so that particular effort has been devoted into developing methods able to estimate it (semi-)automatically. In Haro et al. (2000), the *Translated Poisson Mixture Model* (TPMM) is used to estimate the dimensionality of data in local neighborhoods and partition it accordingly, while *Hidalgo* (Allegra et al., 2020) is a Bayesian extension of *TWO-NN* (Facco et al., 2017), where the dimensionality information is recovered for each point based on the distance to its closest neighbor. Other methods consist in evaluating the local covariance matrix within a pre-specified small volume centered on each data point and analyzing its eigen-decomposition (e.g. Mordohai and Medioni, 2005 and Mordohai and Medioni, 2010). Despite all effort spent in recovering low-dimensional noisy manifolds in a noisy environment, to the best of our knowledge, a complete, coherent formulation that is also flexible and straight-forward to use is still missing in the current scenario. In order to address all the issues presented so far in the context of multiple noisy manifolds learning in a noisy environment, we propose a cohesive toolbox for denoising and 1D manifold (filament) extraction. The toolbox consists of five methodologies that have been exhaustively tested in separate works, however this is the first time that they all come to fruition in a single environment.

Table 1

Information about adopted data sets for experimental sections.

Data set	Size	Attributes
Synthetic	4×10^4	\mathbf{t}, h, T_1, T_2
Jellyfish	9×10^4	$\mathbf{t}, \rho, T, [\text{Fe}/\text{H}], m$
Cosmic Web	$\sim 1 \times 10^6$	\mathbf{t}, \mathbf{v}
ω_{Cen}	$\sim 2 \times 10^4$	ℓ, b

The aim of this work is to present all methodologies in detail, highlighting their functionalities and main objectives and demonstrating (with the user in mind) how the various tools can be combined in different ways for a variety of astronomical applications, using both observed and simulated data. We show how their application on astronomical data sets may drive scientific inference on the underlying physical processes and main properties of the studied objects. The complete implementation of all methods can be found at the following online repository: <https://git.lwp.rug.nl/cs.projects/1DREAM>.

1.1. Organization of the paper

In Section 2 we introduce the synthetic data set used for the controlled experiments in the following section. In order to estimate the efficiency of the methodologies, we carefully construct a mock data set to test the success of our tool-kit in fitting and recovering the known properties of the mock. In particular, we create a point-cloud presenting two elongated non-linear filaments. Two variables exhibit well-defined behaviors along and across the two filaments. In Section 3, we present the five different algorithms introduced in this work. We describe each algorithm in detail and apply it to the mock data set. In Section 4 we outline two powerful visualization techniques that take full advantage of the methodologies described in Section 3. Finally, the whole methodology, aided by the visualization techniques, is applied to three different astronomically relevant data sets, namely:

1. A temporal snapshot of a simulated dwarf galaxy falling into the gravitational potential of a Fornax-Cluster-like galaxy cluster. The simulation is performed using a modification to GADGET-2 (Springel, 2005), an N -body/SPH (Smoothed-Particle Hydrodynamics) code, where a moving box follows closely the evolution of the simulated object. Our goal in this case is to study the properties of other simulated quantities, such as temperature and metallicity, to assess if the recovered streams are loci of Star Formation [$\sim 9 \times 10^4$ points].
2. A temporal snapshot of a large scale formation, Dark Matter only, N -body simulation performed with the GADGET-3, where we extract filaments of the cosmic web and study the dynamics of the belonging particles. [$\sim 1.99 \times 10^6$ points]
3. Stellar stream filaments from the GAIA data set. We focus on one particular filament as the remnant of a previous interaction between our Galaxy and an external object [$\sim 2 \times 10^4$ points].

A summary of properties of the individual data sets can be found in Table 1, where the number of particles (Size) and attributes per particle are shown. We describe these data sets and their analysis in detail in Section 5 and draw our conclusions in Section 6.

2. Synthetic data set for controlled experiment

While the methods can be applied without prior knowledge on the data set at hand, it is essential that the results coming

from a carefully constructed case mirror the true nature of the underlying problem. This serves as a synthetic representation for which the ground truth is known to demonstrate our toolbox. The data set described here and used throughout Section 3 is designed to morphologically resemble a jellyfish galaxy by defining three major noisy manifolds \mathcal{M}_k with $k = 1, 2, 3$. \mathcal{M}_1 is a Gaussian distribution having mean $\mu_1 = (0, 0, 0)$ and covariance matrix

$$\Sigma_1 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 26 \end{pmatrix}.$$

Thus manifold \mathcal{M}_1 is a thick, roughly one-dimensional, elongated structure. The thickness of the manifold is large enough to connect manifolds \mathcal{M}_2 and \mathcal{M}_3 . In this example, this manifold represents the “head” of the jellyfish galaxy. Manifolds \mathcal{M}_2 and \mathcal{M}_3 represent two streams departing from the “head” of the jellyfish, simulating the effect of a dynamical process disrupting the main body of the galaxy. They serve as two distinct parts of the “tail” and they are inherently one-dimensional: their underlying true structure is the unit interval $[-1, 1]$ embedded in \mathbb{R}^3 through mapping functions $f_2 : [-1, 1] \rightarrow \mathbb{R}^3$ and $f_3 : [-1, 1] \rightarrow \mathbb{R}^3$. The mapping functions take the form:

$$f_2(\vartheta) = \begin{bmatrix} 10\vartheta + 8 \\ -3 - (2\vartheta + 2)^2 \\ 5 \sin(\pi\vartheta) + 5 \end{bmatrix}, \quad f_3(\vartheta) = \begin{bmatrix} -10\vartheta - 12 \\ 7 - 4 \sin^2\left(\frac{\pi\vartheta}{2}\right) \\ 5 \cos(\pi\vartheta) - 3 \end{bmatrix} \quad (1)$$

where $\vartheta \in [-1, 1]$. The underlying one-dimensional structure of the manifolds is then convolved with noise in order to obtain a morphological analogy with the structures generally found in astronomical simulations and observations. Both manifolds have an overlapping double noise structure: the thin and dense, inner region (*core*) and the thick, sparser layer (*sparse*), both defined as Gaussian mixtures. The dense core of manifold \mathcal{M}_2 is defined as a flat Gaussian mixture whose $N_2^c = 43$ centers lie on $L_2^c = f_2([-1, 1])$ (each distancing from the adjacent by 0.5). The shared covariance matrix for the core Gaussian mixture is the identity matrix $\Sigma_2^c = \mathbf{I}$ and the component weights are $\pi_2^c = 1/N_2^c$. The sparse Gaussian mixture has $N_2^s = 22$ centers lying on the segment $L_2^s = L_2^c$ (the distance between adjacent centers is 1). The shared covariance matrix is $\Sigma_2^s = 9\mathbf{I}$ and components weights are $\pi_2^s = 1/N_2^s$.

Manifold \mathcal{M}_3 has the same overall structure as manifold \mathcal{M}_2 . The core structure has $N_3^c = 21$ centers, regularly sampling segment $L_3^c = f_3([-1, 1])$, while the sparse structure has $N_3^s = 11$ centers on $L_3^s = L_3^c$. Consequently, the mixture components for core and sparse structures are $\pi_3^c = 1/N_3^c$ and $\pi_3^s = 1/N_3^s$ respectively. The shared covariance matrices are $\Sigma_3^c = \Sigma_2^c$ and $\Sigma_3^s = \Sigma_2^s$ for the core and sparse components, respectively. From each manifold \mathcal{M}_k , $k = 1, 2, 3$ we can now generate a point cloud in \mathbb{R}^3 , obtaining $\mathcal{P}_1, \mathcal{P}_2$ and \mathcal{P}_3 . The union $\mathcal{Q} = \bigcup_{k=1,2,3} \mathcal{P}_k$ of all point clouds represents the morphology of the synthetic data set depicted in Fig. 1. Each manifold and the noisy point distribution are sampled by 10^4 points, making the size of the synthetic data set 4×10^4 points.

2.1. Behavior of simulated physical properties

Following the SPH formulation (Price, 2012, Cossins, 2010 and citations therein), we consider each particle \mathbf{t}_i in data set \mathcal{Q} to sample a spherical volume of radius $r = h_i$. Here h_i is the *smoothing length* equal to the distance of particle \mathbf{t}_i to its 50-th neighbor in data set \mathcal{Q} . Under the assumption of mass preservation (Gingold and Monaghan, 1977), we assign a constant value $m_i = 1$ to the mass contained in each volume sampled by

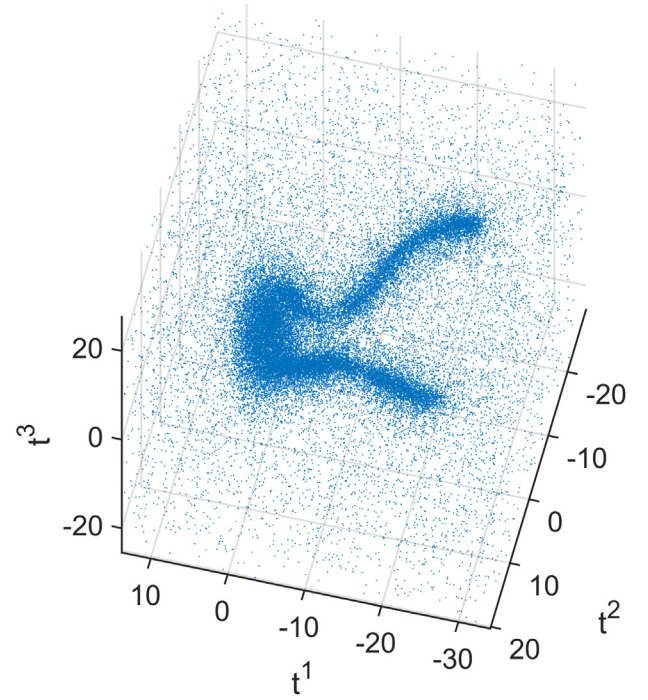


Fig. 1. Final noisy mock data set presenting two noisy elongated filamentary structures, connected to a noisy head and embedded in a noisy environment.

particle \mathbf{t}_i . We can now define the density of the sampled volume as:

$$\rho_i = \frac{m_i}{(4/3)\pi h_i^3}. \quad (2)$$

We also define two additional quantities having particular pre-designed behaviors in proximity to the two manifolds in the data set. Quantity T_1 is defined to be uniformly distributed in the interval $[T_1^{\min}, T_1^{\min} + 1]$ for particles of manifold \mathcal{M}_1 , decreasing from the center of manifold \mathcal{M}_2 ($f_2([-1, 1])$) and sinusoidally varying depending on the distance to the center (radial) of manifold \mathcal{M}_3 ($f_3([-1, 1])$):

$$T_1(\mathcal{P}_1) = X_1 \sim \mathcal{U}(T_1^{\min}, T_1^{\min} + 1); \quad (3)$$

$$T_1(\mathcal{P}_2) = \frac{\delta_{(2,r)}^+ - d(\mathbf{t}_i, L_2^c)_{\mathcal{P}_2}}{4} T_1^{\max}, \quad (4)$$

$$T_1(\mathcal{P}_3) = 1.5 T_{\max} \left\{ 1 + \sin \left[16\pi \frac{d(\mathbf{t}_i, L_3^c)_{\mathcal{P}_3} - \delta_{(3,r)}^-}{\delta_{(3,r)}^+ - \delta_{(3,r)}^-} \right] \right\}. \quad (5)$$

Where $d(\mathbf{t}_i, L_k^c)_{\mathcal{P}_k}$ is the distance between $\mathbf{t}_i \in \mathcal{P}_k$ and the segment $L_k^c = L_k^s = f_k([-1, 1])$, for $j = 2, 3$; $\delta_{(k,r)}^+ = \max_{\mathbf{t}_i \in \mathcal{P}_k} d(\mathbf{t}_i, L_k^c)$ and $\delta_{(k,r)}^- = \min_{\mathbf{t}_i \in \mathcal{P}_k} d(\mathbf{t}_i, L_k^c)$ are the maximum and minimum radial distances, respectively, within all points $\mathbf{t}_i \in \mathcal{P}_k$ from the core of manifold \mathcal{M}_k . Fig. 2a shows the radial profiles of quantity T_1 for the three manifolds.

Quantity T_2 is uniformly sampled within the interval $[T_2^{\min}, T_2^{\min} + 1]$ for all particles in \mathcal{P}_1 , it has an increasing radial profile from the center of manifold \mathcal{M}_2 and it decreases along the longitudinal axis of manifold \mathcal{M}_3 :

$$T_2(\mathcal{P}_1) = X_2 \sim \mathcal{U}(T_2^{\min}, T_2^{\min} + 1); \quad (6)$$

$$T_2(\mathcal{P}_2) = \frac{d(\mathbf{t}_i, L_2^c)_{\mathcal{P}_2} - \delta_{(2,r)}^-}{4} T_2^{\max}, \quad (7)$$

$$T_2(\mathcal{P}_3) = 3 T_2^{\max} \left[\frac{\delta_{(3,L)}^+ - d(\mathbf{t}_i, f_3(-1))}{\delta_{(3,L)}^+ - \delta_{(3,L)}^-} \right], \quad (8)$$

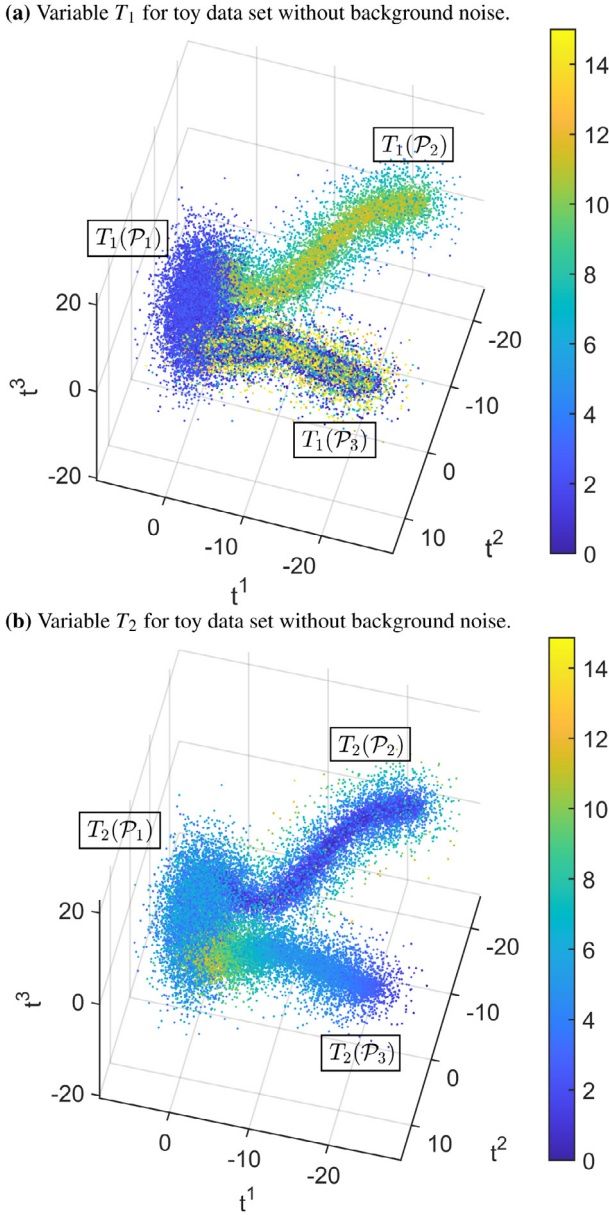


Fig. 2. Distribution of variables T_1 and T_2 for manifolds \mathcal{M}_1 , \mathcal{M}_2 and \mathcal{M}_3 . Panels (a) and (b) mimic a stage of a simulated jellyfish galaxy. Variable T_1 shows a sinusoidal behavior across the radial direction of manifold \mathcal{M}_3 and a decreasing one along the radial direction of \mathcal{M}_2 (panel (a), lower and upper filaments respectively), while being uniformly distributed throughout \mathcal{M}_1 . Variable T_2 is decreasing along the longitudinal direction of manifold \mathcal{M}_3 and the radial direction of manifold \mathcal{M}_2 (panel (b)). It is again uniformly distributed on \mathcal{M}_1 . The uniform noise shown in Fig. 1 is here omitted for clarity.

where the quantities $\delta_{3,L}^+ = \max_{\mathbf{t}_i \in \mathcal{P}_3} d(\mathbf{t}_i, f_3(-1))$ and $\delta_{3,L}^- = \min_{\mathbf{t}_i \in \mathcal{P}_3} d(\mathbf{t}_i, f_3(-1))$ are the maximum and minimum (longitudinal) distances, respectively, within all particles $\mathbf{t}_i \in \mathcal{P}_3$ from point $f_3(-1)$, representing the head of manifold \mathcal{M}_3 . In Fig. 2b the radial profiles of quantity T_2 for the three manifolds can be found. We omit here the background noise.

3. Algorithms

The different methodologies forming the algorithm are here briefly described in order to let the reader have a complete

qualitative overview. Each method is further described in detail in the following sections and graphically depicted in Fig. 3.

LAAT: (Fig. 3a) Taking advantage of the Ant Colony Optimization methodology, *Locally Aligned Ant Technique* (LAAT) aims at enhancing the contrast between high and low density regions in a point cloud via the use of pheromone. This scalar field is used to distinguish between high and low density regions in the data set. By selecting a threshold in the pheromone value, it is possible to filter out particles while preserving those lying in a dense environment. If sub-structures are hidden within the point-cloud, LAAT helps in uncovering them with an adjustable parameter (threshold);

EM3A: (Fig. 3b) The *Evolutionary Manifold Alignment Aware Agents* (EM3 A) algorithm aims at enhancing density contrasts in the data set by pushing particles in high density regions towards an empirically estimated mean curve of the hidden sub-structures, using a similar framework as LAAT. The result of this procedure is a “diffused” point-cloud where the transverse noise to sub-structures is greatly reduced, enabling a more efficient application of the subsequent methodologies;

Dimensionality Index: (Fig. 3c) The resulting points in the diffused data set may belong to structures of different, low, intrinsic dimension. By eigen-decomposition of local neighborhoods, the *Dimensionality Index* assigns to each point in the diffused (and respectively, the noisy) data set, an integer label denoting its intrinsic dimension. The labels are used to partition both the noisy and the diffused data sets into their low-dimensional counterparts;

1D Multi-Manifold Crawling: (Fig. 3d) operates on the one-dimensional partition of the diffused data set. As an iterative procedure, it discovers filaments in the point-cloud and constructs their corresponding skeletons in the data space, while building their low-dimensional representations. The procedure operates a small agent walking (crawling) along the diffused point-cloud following the direction given by the local tangent space estimation. The procedure ends by depletion of the data set of the visited regions by crawling. Its end result is an atlas of structures recovered from the data set, each one with its low- and high-dimensional representations. The previously described methodologies are unable to distinguish between different substructures. Their aim is to enhance the possibility of their detection by filtering out (LAAT) or reduce (EM3 A) background and transverse noise respectively. It is only with 1D Multi-Manifold Crawling that the low-dimensional structures are detected, separated and pre-modeled (via their low-dimensional representation). While the outcome of the previous methodologies is a global point-cloud, the result of 1D Multi-Manifold Crawling is a set of partitions of the data set, where each partition is a detected structure that carries a low-dimensional representation.

Stream GTM: (Fig. 3e) Since ultimately all structures are initially noisy and only by pre-processing we are able to recover their skeleton, *Stream Generative Topographic Mapping* (Stream GTM) builds a probabilistic model for each extracted sub-structure, describing the transverse noise distribution along the manifold itself as a constrained Gaussian mixture model. This allows for a more natural representation of the filaments and serves as a tool for further analyzing the recovered structures.

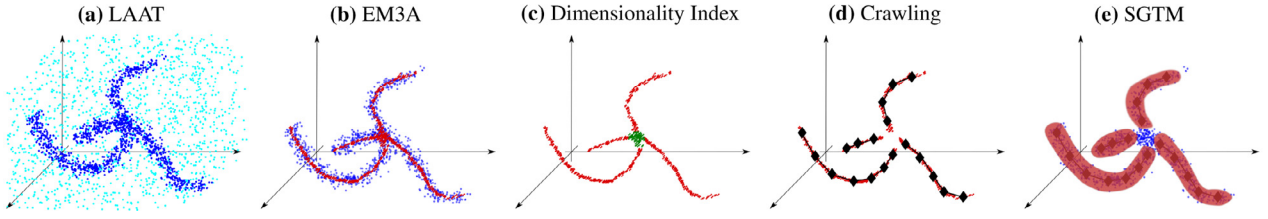


Fig. 3. Sketches depicting the five methodologies for structure detection (a), denoising (b), dimensionality index (c), crawling (d) and modeling (e) proposed in our toolbox.

Via the use of the methodologies composing 1-DREAM it is then possible to identify varying density, low-dimensional regions in any particle data set and to recover filament-like structures hidden within a noisy environment. Furthermore, the structure of individual filaments is regularized via their probabilistic formulation introduced in SGTM. Transverse noise along a detected manifold is here used to achieve smoothness of the manifold structure by modeling it as a mean curve plus noise. While the comparison with other methodologies is not the focus of this work (and will be presented in an upcoming paper¹), 1-DREAM has in this regularization property and advantage with respect to its competitors. The structures recovered via 1-DREAM are indeed more robust to noise, even when taking the stochasticity of the methodologies into account.

3.1. Noise attenuation

We first describe the methodologies developed for the reduction of background (LAAT) and transverse (EM3 A) noise. The pheromone value recovered by LAAT informs us about the background noise level. By thresholding it at specific values (dependent on the data set) the underlying coherent structures emerge from the noisy environment. Filtering out the points with low pheromone value, we obtain the data set to be fed to EM3 A. This methodology reduces transverse noise on the manifolds by pushing points towards their spine (mean curve/surface). These steps are often necessary for further analysis of the hidden structures.

3.1.1. LAAT: Locally Aligned Ant Technique

The Locally Aligned Ant Technique (Taghribi et al., 2022) aims to extract manifolds from noisy data sets in reliance on the idea of Ant Colony Optimization (ACO; Dorigo and Stützle, 2004). The latter is a computational method used for revealing the shortest path between two given data points when multiple routes between them are possible. In this section, we clarify the methodology of LAAT that links between the natural behavior of ants as they follow their own pheromone trails in search of the shortest path to food and the capturing of points that are locally aligned with the major directions of a given manifold. For a complete view of LAAT, we refer the reader to Algorithm 1 in Taghribi et al. (2022).

Consider a data set $\mathcal{Q} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$ consisting of n points such that $\mathbf{t}_i \in \mathbb{R}^D$, then there exists D principle components in a spherical neighborhood $\mathcal{N}_r^i := \mathcal{B}(\mathbf{t}_i, r)$ of radius r around a point \mathbf{t}_i . We denote \mathbf{v}_d and λ_d the local eigenvectors and the corresponding ordered eigenvalues respectively with $d = 1, 2, \dots, D$. That being introduced, LAAT then consists of a random walk in which artificial “ant” jump from a point belonging to the data set to the next where high preference is given to: jumps along the dominant eigenvectors, and paths where an amount of artificially deposited pheromone is accumulated (Dorigo and

Stützle, 2004). Given a path $(\mathbf{t}_j - \mathbf{t}_i)$ between points i and j , the relative normalized weighting of the alignment of this path with a local eigenvector \mathbf{v}_d can be given as follows:

$$w_d^{(i,j)} = \frac{|\cos \alpha_d^{(i,j)}|}{\sum_{d'=1}^D |\cos \alpha_{d'}^{(i,j)}|}. \quad (9)$$

Where $\alpha_d^{(i,j)}$ is the angle between $(\mathbf{t}_j - \mathbf{t}_i)$ and \mathbf{v}_d . Furthermore, the normalized eigenvalues show the relative importance of the different eigenvectors, and are given by the following:

$$\tilde{\lambda}_d^{(i)} = \frac{\lambda_d^{(i)}}{\sum_{d'=1}^D \lambda_{d'}^{(i)}} \quad (10)$$

This then allows us to define the preference of the jump from \mathbf{t}_i to \mathbf{t}_j that is aligned with the local eigen-directions. The preference is given by:

$$E^{(i,j)} = \sum_{d=1}^D w_d^{(i,j)} \cdot \tilde{\lambda}_d^{(i)} \quad (11)$$

Preferences are normalized ($\tilde{E}^{(i,j)}$) so that they sum to 1 within the neighborhood \mathcal{N}_r^i . Moreover, by defining an amount of pheromone $F^j(t)$ for a point \mathbf{t}_j at a time t , the above preference will allow for the accumulation of pheromone on the points aligning with the manifold. As inspired by nature, an evaporation rate $0 < \zeta < 1$ is incorporated in the definition of the pheromone thus serving the purpose of decreasing its amount on less visited points. The pheromone quantity is:

$$F^j(\tau + 1) = (1 - \zeta) \cdot F^j(\tau), \quad (12)$$

Again, pheromone quantities are normalized ($\tilde{F}^j(\tau)$) so that they sum to 1 within \mathcal{N}_r^i . Combining Eqs. (11) and (12) allows us to define the total preference of the jump from \mathbf{t}_i to \mathbf{t}_j :

$$V^{(i,j)}(\tau) = (1 - \kappa) \tilde{F}^j(\tau) + \kappa \tilde{E}^{(i,j)}. \quad (13)$$

Where $\kappa \in [0, 1]$ is a parameter which adjusts the relative importance of the two terms. Finally, the jump probabilities can be defined as:

$$P(j|i, \tau) = \frac{\exp(\beta V^{(i,j)}(\tau))}{\sum_{j' \in \mathcal{N}_r^i} \exp(\beta V^{(i,j')}(\tau))}, \quad (14)$$

Here, $\beta > 0$ is a parameter that is analogous to the inverse of temperature in statistical physics (Taghribi et al., 2022). Afterwards, choosing a set of hyper-parameters given by the number of ants N_{ants} , epochs N_{epoch} , and steps of each ant N_{steps} is necessary for the completion of the random walk. A random starting point is chosen such that the random walk begins from the denser neighborhoods. In practice, this means that given the median $\tilde{\mathcal{N}}$ of the set of neighborhoods $\mathcal{N} = \{|\mathcal{N}_r^{(i)}| \mid \mathbf{x}_i \in \mathcal{Q}\}$, the condition for a random starting point i is:

$$|\mathcal{N}_r^{(i)}| \geq \tilde{\mathcal{N}} \quad (15)$$

In a given epoch, the ants will then perform the random walk on the points in \mathcal{Q} for N_{steps} and with the jump probabilities

¹ This additional work can be found at <https://git.lwp.rug.nl/cs.projects/1DREAM>.

Table 2
Full list of parameters for LAAT.

$r \in \mathbb{R}$	Neighborhood radius
$\zeta \in \mathbb{R}$ ($\zeta = 0.1$)	Evaporation rate
$\kappa \in \mathbb{R}$ ($\kappa = 0.5$)	Shape v , Pheromone
$\beta \in \mathbb{R}$ ($\beta = 10$)	Inverse temperature
$\gamma \in \mathbb{R}$ ($\gamma = 0.05$)	Deposited pheromone
$F_{th}^j \in \mathbb{R}$	Pheromone threshold
$\tilde{N} \in \mathbb{N}$ ($\tilde{N} = 5$)	Neighborhood threshold
$N_{epoch} \in \mathbb{N}$ ($N_{epoch} = 10$)	Epochs
$N_{steps} \in \mathbb{N}$ ($N_{steps} = 2500$)	Steps per epoch
$N_{ants} \in \mathbb{N}$ ($N_{ants} = 500$)	Ants

defined in (14). To update the value of the pheromone quantity, the indices of the points visited by a given ant ℓ is stored in a route multi-set A^ℓ which allows us to count the multiplicity of visits to the points in the data set. The value of the pheromone quantity on a given point j is updated according to the following formula:

$$F^j(\tau) = F^j(\tau - 1) + v(j)\gamma \quad \forall j \in A^\ell, \quad (16)$$

where γ is a constant value denoting the amount of pheromone deposited, and $v(j)$ is the multiplicity of element j in A^ℓ . Therefore, with the enforced pheromone evaporation rate and the defined jump probabilities, the pheromone will accumulate along the points aligned with given manifolds, and will dissipate in more scattered regions, hence highlighting the structures in a noisy data set. We refer the reader to Taghribi et al. (2022) for a demonstration of the high robustness of results to changes in the previously defined parameters, and a comparison with state-of-the-art methods of similar purpose. The application of the LAAT methodology to the synthetic data set described previously is shown in Fig. 4. For this analysis we used a radius $r = 2$ and a neighborhood size threshold of $\tilde{N} = 5$. The blue inner structure is revealed from within the whole noisy data set. While visually, the same structure can be identified in Fig. 1, LAAT introduces a scalar field on the data set as a proxy of relevant dense regions. It is then straightforward to provide a pheromone threshold (e.g. in our case $F_{th}^j = 20$) that enables filtering of the noisy data set. The filtered data set will contain only the most relevant regions. A complete list of all parameters in the LAAT methodology is given in Table 2. Free-parameters are denoted by a $*$ symbol and their values specified in each experimental study (Sections 5.1–5.3). The values suggested for the rest of the parameters are the ones used throughout this work.

3.1.2. EM3A: Evolutionary manifold alignment aware agents

While LAAT isolates points in proximity to high-density regions, ideally the true manifold lies within the noisy point-cloud that samples it. In order to find the mean curve of manifolds, by reducing the amount of transverse noise, a secondary step in this analysis is necessary. Intuitively, we would like to be able to push points in proximity to over-dense regions in the data space, towards the unknown mean curve of the manifold. We propose a procedure, devoted to approximate the true nature of manifolds by pulling nearby points towards an empirically estimated, local mean curve. This solution is also inspired by ant colony behavior for picking up and dropping carried objects. This section explains the steps followed by EM3 A, a method introduced in Mohammedi and Bunte (2020) to employ this behavior by instructing agents walking through the data space to pick up the data points and place them in a closer proximity to the manifolds. Since the net effect of the methodology is to move points towards high density, low-dimensional regions, the methodology shares similarities with the work presented in Wu et al. (2018), where the task is modeled as a diffusion process. For this reason we will often refer to the data set obtained via EM3 A as a “diffused” one.

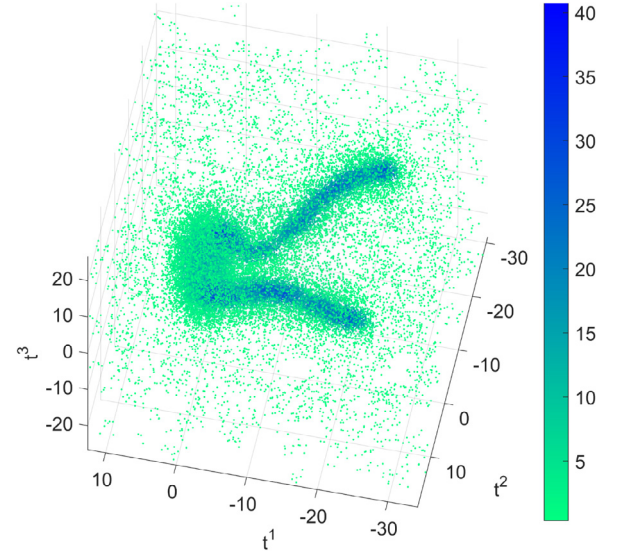


Fig. 4. Pheromone value for every point in the data synthetic data set emulating a Jellyfish galaxy at the end of the procedure.

The first step in accomplishing the above is to define an approximation strategy to recognizing the manifold structure. In that pursuit, local PCA is employed again to estimate the local eigenvalues and eigenvectors within neighborhoods of radius r centered on the data points. Given data set \mathcal{Q} and $\mathbf{t}_i \in \mathcal{Q}$, the eigenvectors \mathbf{v}_d^i and normalized eigenvalues $\tilde{\lambda}_d^i$ for the neighborhood $\mathcal{N}_r^i := \mathcal{B}(\mathbf{t}_i, r) \cap \mathcal{Q}$. Defining the saliencies (Mordohai and Medioni, 2010) $S_i = i \times (\tilde{\lambda}_d^{(i)} - \tilde{\lambda}_d^{(i+1)})$ as the expansion coefficient for the local covariance matrix, the intrinsic dimensionality of the manifold is estimated as given in Wang et al. (2008), Mordohai and Medioni (2005) by:

$$\hat{d} = \arg \max_i S_i. \quad (17)$$

The set $\{\mathbf{v}_1^i, \dots, \mathbf{v}_{\hat{d}}^i\}$ can then be used as an approximation to the tangent space of manifold \mathcal{M} at \mathbf{t}_i .

Next, a random walk is initiated in which agents are reinforced to move data points closer to the underlying manifolds in the data set. Let \mathbf{U} be the matrix whose columns are the first d eigenvectors of \mathcal{N}_r^i , and let $\boldsymbol{\mu}$ be the kernel average of \mathbf{t}_i 's neighbors, then the distance to a manifold \mathcal{M} is estimated by:

$$\delta^{\mathcal{M}}(i) = \|(I - \mathbf{U}\mathbf{U}^T)(\boldsymbol{\mu} - \mathbf{t}_i)\|. \quad (18)$$

Here $\|\cdot\|$ is the Euclidean norm. We now define the weights and probabilities associated with the random walk. For each point \mathbf{t}_j within \mathcal{N}_r^i , the following weights are defined:

$$w(\mathbf{t}_i, \mathbf{t}_j) = \begin{cases} 1 - \frac{\delta^{\mathcal{M}}(i)}{\iota} & \delta^{\mathcal{M}}(i) \leq \iota \\ 0 & \delta^{\mathcal{M}}(i) > \iota. \end{cases} \quad (19)$$

The parameter ι is chosen such that 50% of neighbors have none-zero weights. The agent jump probability to the next destination is then given by:

$$P(\mathbf{t}_i, \mathbf{t}_j) = \frac{w(\mathbf{t}_i, \mathbf{t}_j)}{\sum_{\mathbf{m} \in \mathcal{N}_r^i} w(\mathbf{t}_i, \mathbf{t}_m)}. \quad (20)$$

Having defined this jump probability, the agents are encouraged to remain close to the manifold. Since the agents are not only walking, but also picking up and dropping down points, what is then required is to define a pick-up probability for the agent of

data point \mathbf{t}_j :

$$P_{\text{pick}}(\mathbf{t}_j) = \frac{1 - w(\mathbf{t}_i, \mathbf{t}_j)}{\sum_{m \in \mathcal{N}_i^r} (1 - w(\mathbf{t}_i, \mathbf{t}_m))}. \quad (21)$$

In other words, the probability to be picked up increases with the distance of the point from the tangent space. Since we aim at enhancing the density contrast between points more likely belonging to the manifold and all others, at each time t we move these points towards the manifold along the complement of the tangent space. The displacement update reads:

$$\mathbf{t}_j^{\text{new}} = \mathbf{t}_j^{\text{old}} + \eta(I - UU^T)(\boldsymbol{\mu} - \mathbf{t}_i^{\text{old}}). \quad (22)$$

Here $\eta > 0$ is the learning rate controlling the amount of displacement produced. In addition to the denoising, as mentioned previously, the topological nature of the manifold should also be preserved under the above steps. To ensure that, over-smoothing of the manifold is avoided by introducing a threshold such that the agents can only change the neighborhood if the mean distance of neighbors to the tangent space is larger than the threshold.

The performance of the above method is clearly dependent on the chosen radius of the neighborhood (Kaslovsky and Meyer, 2014). Since highly curved manifolds require smaller radii while the suppression effect of the noise requires larger radii, it is difficult to choose a proper value for r without prior knowledge of the manifold properties. Therefore, as a solution to this problem, EM3 A combines the above procedure with Evolutionary Game Theory (EGT) concepts to automatically adapt the radius parameter.

Since r is a continuous variable, we assume that it lies within the range $[R_{\min}, R_{\max}]$ discretized into m smaller intervals. To link this step to EGT, each interval is therefore viewed as an evolutionary strategy within a population of m strategies. Taking p_1, \dots, p_m as the frequencies of each strategy, we let $\mathbf{p} = \{p_1, \dots, p_m\}$ denote the distribution of strategies within the population. In a generation t , each agent randomly selects a strategy while following a population share distribution $\mathbf{p}^{(t)} = [p_\ell^{(t)}]_{m \times 1}$, where $p_\ell^{(t)}$ is the population share (frequency) of the ℓ -th strategy at the current generation. For this agent with a strategy ℓ , the neighborhood radius is uniformly selected from the interval $[r_\ell, r_{\ell+1}]$.

Next, a copy of the data set is provided for all the agents on which they perform the random walk with N_s steps according to the above described rules. The output of each walk is averaged to form the updated data set and the fitness of each strategy is computed. Letting S_ℓ denote the number of agents with strategy ℓ and N_ℓ the number of times they change the data set at a given generation, then the fitness f_ℓ of strategy ℓ follows:

$$f_\ell = \frac{N_\ell}{S_\ell \cdot N_s}. \quad (23)$$

For a given strategy ℓ , the rate of change of its frequency \dot{p}_ℓ/p_ℓ , measures its evolutionary success. This measure can be equated to the difference between the fitness of the strategy and the average fitness of the population. In other words for a generation t we can define:

$$\frac{\dot{p}_\ell}{p_\ell} = f_\ell - \bar{f}. \quad (24)$$

The above equation can be rewritten in the following iterative form:

$$p_\ell^{(t+1)} = p_\ell^t + p_\ell^t(f_\ell^t - \bar{f}^t). \quad (25)$$

Therefore, in the following generations, the agents will choose their strategy from the distribution $\mathbf{p}^{(t+1)} = [p_\ell^{(t+1)}]$. Iterating for a given number of generations will thus lead to the denoising of the manifolds in a noisy data set while maintaining the properties

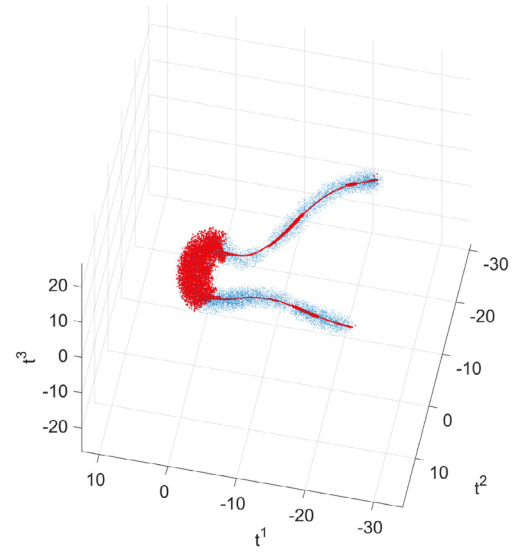


Fig. 5. Data set of points selected s.t. $F(t = t_{\text{end}}) \geq F_{\text{Th}}$ (blue dots) and diffused data set obtained with EM3 A (red dots).

Table 3

Full list of parameters for EM3A.

$R_{\min} \in \mathbb{R}$	Minimum radius
$R_{\max} \in \mathbb{R}$	Maximum radius
$\iota \in \mathbb{R} (\iota > 0)$	Jump weight (adaptive)
$\eta \in \mathbb{R} (\eta = 1e-2)$	Learning rate
$N_s \in \mathbb{N} (N_s = 500)$	Number of steps
$N_g \in \mathbb{N} (N_g = 5)$	Number of generations

of the embedded structures. Results of EM3 A applied to the synthetic data filtered according to the pheromone value are shown in Fig. 5 (red dots). The processed data is sensibly denser along the mean curve of the two elongated manifolds, while it is roughly unchanged for manifold \mathcal{M}_1 (sampled by point-cloud \mathcal{P}_1). The over-densities in the middle of the two manifolds (close to the high-curvature regions) are likely to be caused by slight variations in the pheromone value recovered by LAAT. The crisp selection of particles with high pheromone may have caused varying point-density along the manifolds, and this is converted by EM3 A in a stronger push towards the over-dense regions. Indeed, the denser red regions are always delimited by gaps (under-dense regions) in the filtered point-cloud (blue dots). All parameters of EM3 A are listed in Table 3. The free parameters used for the synthetic data set are $R_{\min} = 1$ and $R_{\max} = 2$.

3.2. Modeling

In the following we describe the methodologies aimed at identifying the local intrinsic dimensionality of structures in the data set (Dimensionality Index) on a point-by-point basis (Section 3.2.1). The one-dimensional points are used for detection of filaments via Crawling (Section 3.2.2) and its results as initialization for SGTM 3.2.3. Since these methodologies aim at detecting and recovering individual sub-structures in the data set, we refer to this step as “modeling”.

3.2.1. Dimensionality index

While in first approximation, the approach used in Section 3.1.1 is enough to roughly estimate the local dimensionality of a neighborhood, we dedicate this section to a more refined version of dimensionality index. In light of the new data

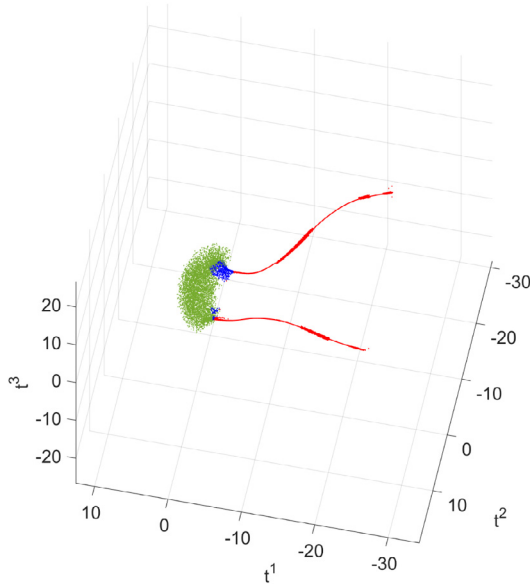


Fig. 6. Partition of data sets $\tilde{\mathcal{Q}}$ into its corresponding one, two and three-dimensional subsets: $\tilde{\mathcal{Q}}_1$ (red) $\tilde{\mathcal{Q}}_2$ (blue) and $\tilde{\mathcal{Q}}_3$ (green).

set obtained via LAAT and EM3 A, the new estimate of local dimensionality incorporates continuity information about the local structure, previously hidden by noise. Around each $\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$ we perform local PCA using points from $\mathcal{N}_r^i = \mathcal{B}(\tilde{\mathbf{t}}_i; r) \cap \tilde{\mathcal{Q}}$, obtaining eigenspectrum $\lambda_{i,1} \geq \lambda_{i,2} \geq \dots \geq \lambda_{i,d}$.

The dimensionality index Δ_i^0 of $\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$ used in Wang et al. (2008) (limited to 3-dimensional data) is obtained as in Eq. (17). However, a more accurate dimensionality index can be found in Canducci et al. (2022). We summarize it in the current section. The normalized eigen-spectrum $\tilde{\Lambda}_i$ (see Eq. (10)) of each point $\tilde{\mathbf{t}}_i$'s neighborhood is mapped onto the Simplex of multinomial distributions. The geodesic distance of each Λ_i with respect to the vertices $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is evaluated on the simplex, where $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (1/2, 1/2, 0)$, $\mathbf{e}_3 = (1/3, 1/3, 1/3)$ represent the eigen-spectra of ideal 1-, 2- and 3-dimensional neighborhoods respectively. Then, the dimensionality index of point $\tilde{\mathbf{t}}_i$ is the index j corresponding to the closest vertex, under the geodesic distance $d_j(\Lambda_i, \mathbf{e}_j)$:

$$\Delta_i^G = \arg \min_j d_j(\tilde{\Lambda}_i, \mathbf{e}_j), \quad (26)$$

$$d_j(\tilde{\Lambda}_\ell, \tilde{\Lambda}_m) = 2 \arccos \left(\sum_{k=1}^D \sqrt{\tilde{\Lambda}_\ell^k \cdot \tilde{\Lambda}_m^k} \right). \quad (27)$$

We also propose a “soft” version of dimensionality index, by imposing a kernel $K(\tilde{\Lambda}_i; \mathbf{e}_j)$ on each prototypical vertex of the Simplex. We chose to use a Gaussian smoothing kernel s.t.:

$$K(\tilde{\Lambda}; \mathbf{e}_j) = \exp \left[-\frac{d_j(\tilde{\Lambda}, \mathbf{e}_j)^2}{2s^2} \right], \quad (28)$$

where s is the geodesic distance on the simplex between any vertex and the equidistant point on the Simplex with respect to all vertices. This kernelization of the geodesic distances on the Simplex imposes a distribution:

$$P_i(j) = \frac{K(\tilde{\Lambda}_i; \mathbf{e}_j)}{\sum_{k=1}^D K(\tilde{\Lambda}_i; \mathbf{e}_k)}. \quad (29)$$

In order to take into account the smoothness of manifolds in data space, we impose a smoothing kernel in that space on each point

$\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$ s.t.:

$$c(i, l) = \exp \left[-\|\tilde{\mathbf{t}}_i - \tilde{\mathbf{t}}_l\|^2 / (2r^2) \right].$$

The smoothed normalized index distribution reads:

$$P_i^S(j) = \frac{1}{\sum_{\tilde{\mathbf{t}}_l \in \mathcal{B}(\tilde{\mathbf{t}}_i, r)} c(i, l)} \sum_{\tilde{\mathbf{t}}_l \in \mathcal{B}(\tilde{\mathbf{t}}_i, r)} c(i, l) \cdot P_l(j), \quad (30)$$

where the sum is taken over diffused points $\tilde{\mathbf{t}}_l$ in the spherical neighborhood of $\tilde{\mathbf{t}}_i$ of radius r . The smoothed dimensionality index of $\tilde{\mathbf{t}}_i$ is then

$$\Delta_i^S = \arg \max_j P_i^S(j). \quad (31)$$

Every point in $\tilde{\mathcal{Q}}$ (and its noisy counterpart \mathcal{Q}) can be assigned to the respective d -dimensional subset $\tilde{\mathcal{Q}}_d$ (\mathcal{Q}_d), creating a partition of the original set into:

$$\tilde{\mathcal{Q}}_d = \{\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}} \mid \Delta_i^S = d\} \quad (32)$$

$$\mathcal{Q}_d = \{\mathbf{t}_i \in \mathcal{Q} \mid \Delta_i^S = d\}, \quad (33)$$

such that $\bigcup_{d=1}^D \mathcal{Q}_d = \mathcal{Q}$. The results of the application of dimensionality index to the point cloud defined in Section 2 are shown in Fig. 6 and were obtained with the same radius used for LAAT ($r = 2$). For a better visualization of the results we only show here the diffused points as divided into 1-, 2-, 3-dimensional sets, represented in red, blue and green color respectively (\mathcal{Q}_1 , \mathcal{Q}_2 and \mathcal{Q}_3 respectively). Although the synthetic data set described in Section 2 does not contain intrinsically two-dimensional points, the dimensionality index proposed in Eq. (31) recovers $\tilde{\mathcal{Q}}_2 \neq \emptyset$. This discrepancy can be attributed to the EM3 A algorithm diffusing points too strongly towards high-density regions. However, since we are interested in detecting the one-dimensional structures (streams) in the data set, upon visual inspection, the recovered estimation of $\tilde{\mathcal{Q}}_1$ and \mathcal{Q}_1 (red points in Fig. 6) is acceptable and the contamination of 2-dimensional (blue) points is minimal.

Note that the only parameter required by the dimensionality index is the neighborhood radius r .

3.2.2. Crawling: Multiple manifolds 1D1d crawling

We describe here a recursive algorithm that enables us to separate all distinct one-dimensional manifolds contained in data sets \mathcal{Q}_1 and $\tilde{\mathcal{Q}}_1$ while sampling them in representative sets of points and building their respective low dimensional representations. The end result of this technique (see Fig. 7), as opposed to LAAT (Section 3.1.1), is a discrete skeleton for each one-dimensional manifold in the data set. In fact, LAAT only highlights over-densities within a point-cloud, without partitioning into its low-dimensional components. With Crawling, each skeleton is assumed to lie on a high dimensional embedding of the unit interval (a bent and stretched version of it), sampled by a finite set of points representing the “steps” taken by the agent while walking (crawling) on $\tilde{\mathcal{Q}}_1$. The results from this algorithm are used to initialize Stream GTM through a parametric mapping function $\mathbf{f} : [-1; 1] \rightarrow \mathbb{R}^D$ via linear regression applied to the parameters \mathbf{W} . The only main assumption of the algorithm is that at the selected size (given by the radius parameter, in this case $r = 2$), the tangent space to each manifold is isomorphic to \mathbb{R} .

Initialization

We first initialize the residuals set $\mathcal{R} = \tilde{\mathcal{Q}}_1$. This data set is used as a reference for regions that have been visited by crawling, leaving $\tilde{\mathcal{Q}}_1$ unscathed. Initially, a single “seed” $\tilde{\mathbf{t}}_0 \in \tilde{\mathcal{R}}$ is randomly selected and PCA applied to its local neighborhood of radius r : $\mathcal{B}(\tilde{\mathbf{t}}_0, r) \cap \tilde{\mathcal{R}}$. The unit eigen-vector $\hat{\mathbf{v}}_0 = \frac{\mathbf{v}_0}{\|\mathbf{v}_0\|}$, associated to the largest eigen-value λ_1 , spans the tangent space to manifold \mathcal{M}_k at point $\tilde{\mathbf{t}}_0$: $T_{\tilde{\mathbf{t}}_0} \mathcal{M}_k$. Eigen-vector $\hat{\mathbf{v}}_0$, being the direction where most

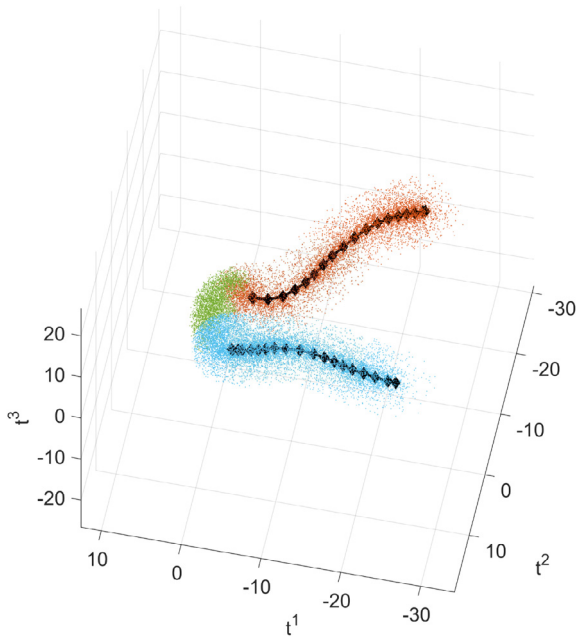


Fig. 7. Extracted skeletons (black diamonds and line) of the two manifolds after crawling, overlay the 1D data set extracted by means of the dimensionality index (red lines). The point-cloud surrounding each skeleton is the corresponding recovered noisy manifold.

of the neighborhood's variance is preserved, gives us the initial preferential direction for crawling. We can now estimate two new points along direction $\hat{\mathbf{v}}_0$ at distance $\beta \cdot r$ from $\tilde{\mathbf{t}}_0$:

$$\mathbf{z}_n^\pm = \tilde{\mathbf{t}}_0 \pm \beta \cdot r \cdot \hat{\mathbf{v}}_0, \quad (34)$$

where $\beta = 0.75$ is a regularization parameter aimed at mitigating the effect of outliers on PCA and index n identifies the iteration number (during the initialization $n = 1$). In order to keep the crawling adherent to manifold \mathcal{M}_k , for every new candidate \mathbf{z}_n^\pm we compute its closest neighbor in data set $\tilde{\mathcal{R}}$:

$$\tilde{\mathbf{t}}_n^\pm = \arg \min_{\tilde{\mathbf{t}} \in \tilde{\mathcal{R}}} (\|\tilde{\mathbf{t}} - \mathbf{z}_n^\pm\|) \quad (35)$$

under the condition that $\|\tilde{\mathbf{t}}_n^\pm - \tilde{\mathbf{t}}_0\| \leq r$. This condition enforces the maximum length between two adjacent points on manifold \mathcal{M}^k to never exceed the neighborhood radius r . After the estimation step, we initialize the set of representative points of manifold \mathcal{M}^k as $\bar{\mathcal{P}}^k = \{\tilde{\mathbf{t}}_0, \tilde{\mathbf{t}}_1^+, \tilde{\mathbf{t}}_1^-\}$ and the low-dimensional counterpart $\mathcal{P}^k = \{0, 1, -1\}$. The direction $\hat{\mathbf{v}}_0$, being the tangent space to \mathcal{M}^k at point $\tilde{\mathbf{t}}_0$, is preserved as a member of the tangent bundle (see Tu, 2010) to manifold \mathcal{M}^k : TM^k . The last step in the initialization phase removes the neighborhood of point $\tilde{\mathbf{t}}_0$ from \mathcal{R} overwriting the set.

Crawling update

After the initialization phase is completed, at every iteration n , crawling is recursively applied to every point identified in the previous iteration $n - 1$ following:

1. Seed selection: From $\bar{\mathcal{P}}^k$ select point $\tilde{\mathbf{t}}_{n-1}^+$ and compute the neighborhood $\mathcal{N} = \mathcal{B}(\tilde{\mathbf{t}}_{n-1}^+, r) \cap \mathcal{R}$. Applying PCA to \mathcal{N} , compute the unit principal component $\hat{\mathbf{u}}_{n-1}$.

2. Parent point recovery: Recover $\tilde{\mathbf{t}}_{n-2}^+$ and $\hat{\mathbf{v}}_{n-2}^+$ the parent point and corresponding tangent vector of $\tilde{\mathbf{t}}_{n-1}^+$. For “parent” we mean the point that generated $\tilde{\mathbf{t}}_{n-1}^+$ in iteration $n - 2$ of crawling.

3. Tangent space projection: Since PCA is a rotationally invariant method, $\hat{\mathbf{u}}_{n-1}$ is not a priori identifiable with the current crawling direction (it could be oriented as its inverse vector). We solve this issue by computing the angle θ between $\hat{\mathbf{u}}_{n-1}$ and $\hat{\mathbf{v}}_{n-2}^+$

$$\theta = \arccos(\hat{\mathbf{u}}_{n-1} \cdot \hat{\mathbf{v}}_{n-2}^+), \quad (36)$$

where (\cdot) denotes the scalar product. The new crawling direction $\hat{\mathbf{v}}_{n-1}^+$ is given by:

$$\hat{\mathbf{v}}_{n-1}^+ = \begin{cases} +\hat{\mathbf{u}}_{n-1} & \text{if } -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ -\hat{\mathbf{u}}_{n-1} & \text{if } \frac{\pi}{2} < \theta < \frac{3\pi}{2} \end{cases}. \quad (37)$$

4. Updates: Using Eqs. (34) and (35), a new point lying on manifold \mathcal{M}^k in direction $\hat{\mathbf{v}}_{n-1}^+$ is found and added to $\bar{\mathcal{P}}^k$. The latent space \mathcal{P}^k and tangent bundle TM^k are also updated coherently. Finally, the neighborhood \mathcal{N} is subtracted from data set \mathcal{R} .

When the condition $\|\tilde{\mathbf{t}} - \mathbf{z}_n^\pm\| \leq r$ does not hold for any $\tilde{\mathbf{t}} \in \mathcal{R}$ an end of the manifold is encountered and crawling is suppressed along the current direction. The same procedure is applied to point $\tilde{\mathbf{t}}_{n-1}^-$, using parent direction $\hat{\mathbf{v}}_{n-2}^-$, until the second end of the manifold is found.

Once both ends of a manifold are detected, we repeat the *Initialization* and *Update* phases on data set \mathcal{R} , which, at the end of iteration k of the procedure contains all points of data set $\tilde{\mathcal{Q}}_1$ except the ones extracted from manifolds up to \mathcal{M}^k . Since data set \mathcal{R} is recursively depleted of points, we expect its size to converge to zero after a certain number of iterations of the whole procedure. This consideration gives us a criterion for halting the crawling algorithm. When $|\mathcal{R}| \leq v$, $v \geq 0$ being a user-specified threshold, assuming the processing of $\tilde{\mathcal{Q}}_1$ took K runs, the procedure results in a collection of extracted one-dimensional manifolds represented by the sets $\{\bar{\mathcal{P}}^k\}_{k=1}^K$, containing sampled points from data set $\tilde{\mathcal{Q}}_1$ representative of manifolds $\{\mathcal{M}^k\}_{k=1}^K$, $\{\mathcal{P}^k\}_{k=1}^K$, the associated low-dimensional counterparts, and $\{TM^k\}_{k=1}^K$, the respective tangent bundles.

Noisy manifolds recovery

Assuming that set $\bar{\mathcal{P}}^k$ has sampled manifold \mathcal{M}^k with L^k points, for every point $\tilde{\mathbf{t}}^\ell \in \bar{\mathcal{P}}^k$ we compute $\mathcal{N}_\ell^k := \mathcal{B}(\tilde{\mathbf{t}}^\ell, r) \cap \mathcal{Q}_1$. We define the noisy sample of manifold \mathcal{M}^k as the union of all neighborhoods of radius r computed on all points $\tilde{\mathbf{t}}_\ell$, $\ell = 1, \dots, L^k$: $\mathcal{A}^k = \bigcup_{\ell=1}^{L^k} \mathcal{N}_\ell^k$. By performing this assignment for every set $\bar{\mathcal{P}}^k$, we obtain the K sets containing samples of the noisy manifolds detected in the different runs of crawling. We now have K manifolds and for each \mathcal{M}^k in \mathcal{Q}_1 , with $k = 1, \dots, K$, three unique sets:

- $\bar{\mathcal{P}}^k$: contains all points sampled from $\tilde{\mathcal{Q}}_1$ at distance of at most r , forming a skeleton for the manifold;
- \mathcal{P}^k : the low-dimensional representation of set $\bar{\mathcal{P}}^k$;
- \mathcal{A}^k : the set of all points describing its noisy structure, sampled from \mathcal{Q}_1 .

We also recover the tangent bundle TM^k associated to manifold \mathcal{M}^k , by collecting all tangent directions to the manifold on points in $\bar{\mathcal{P}}^k$. For completeness, the list of parameters used in Crawling are presented in Table 4.

3.2.3. SGTm: Stream GTM

The Generative Topographic Mapping (GTM, Bishop et al., 1998b) is a generative algorithm used generally for dimensionality reduction and density modeling of high dimensional, noisy

Table 4
Full list of parameters for Crawling.

$r^* \in \mathbb{R}$	Neighborhood radius
$\beta \in \mathbb{R} (\beta = 0.75)$	Jump tolerance

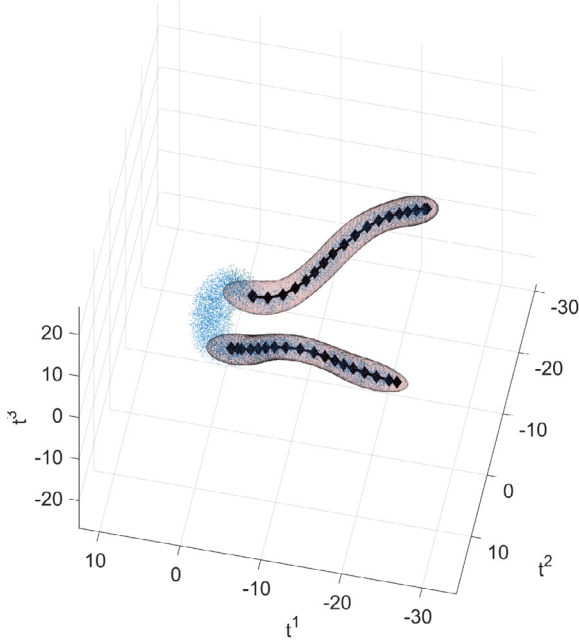


Fig. 8. Iso-surfaces of the Probability Density Function (PDF) of the individual probabilistic models obtained via SGTm for each manifold in the synthetic data set.

data sets. It is the probabilistic formulation of Self-Organizing Map (SOM, Kohonen, 1982) and it aims at modeling low dimensional structures in high-dimensional data sets as constrained Gaussian Mixtures. In its original formulation, the Gaussian centers are constrained to lie on the principal components of the data set, while the noise model is assumed to be spherical. Our formulation differs from the original by imposing a structure on the centers co-linear with the manifold's skeleton and a manifold-aligned noise model (replacing the spherical Gaussian formulation). The proposed methodology is a simplification of the one proposed in Canducci et al. (2022), by noting that non-intersecting one-dimensional graphs can always be embedded on the unit segment of the real line via a non-linear, parametric mapping.

Let us consider manifold \mathcal{M}^k found in data set \mathcal{Q}_1 by crawling on data set $\tilde{\mathcal{Q}}_1$. In the following, we describe Stream GTM applied to a single manifold \mathcal{M}^k , dropping superscript k (for readability reasons) on every component derived by crawling. However, this methodology is performed for every detected \mathcal{M}^k , with $k = 1, \dots, K$.

We first initialize the latent one-dimensional structure of the manifold by scaling set \mathcal{P} , so that it lies on the interval $[-1; 1]$:

$$x_\ell = -1 + \frac{p_\ell - \min(\mathcal{P})}{\max(\mathcal{P}) - \min(\mathcal{P})} \quad \forall p_\ell \in \mathcal{P}. \quad (38)$$

Calling $\mathcal{X} = \{x_\ell, \ell = 1, \dots, L\}$ the scaled \mathcal{P} , we can define a set of S radial basis functions (RBFs) ϕ_1, \dots, ϕ_S , centered on a subset of \mathcal{X} :

$$\phi_s(x_\ell) = \exp \left[-\frac{(x_s - x_\ell)^2}{2\sigma^2} \right]. \quad (39)$$

Table 5
Full list of parameters for SGTm.

$r^* \in \mathbb{R}$	Neighborhood radius
$S^* \in \mathbb{N}$	Number of RBFs
$\nu \in \mathbb{R} (\nu = 1e^3)$	Regularization Cov. matrix
$\omega \in \mathbb{R}$	Scaling factor Cov. matrix (adaptive)

Here σ is computed as the mean distance between adjacent centers: $\sigma = \sum_{s=1}^{S-1} \|x_s - x_{s+1}\| / (S-1)$. The centers x_s of the RBFs are sampled regularly from \mathcal{X} .

The mapping of points from the latent space \mathcal{X} to embedded points in $\bar{\mathcal{P}}$ is achieved by the function $\mathbf{y}(x; \mathbf{W})$, governed by the $S \times D$ matrix of parameters \mathbf{W} . Given the definition of RBFs in Eq. (39) we can define the mapping function as:

$$\mathbf{y}(x; \mathbf{W}) = \Phi(x)\mathbf{W} \quad (40)$$

where \mathbf{x} is the column vector containing points in \mathcal{X} and $\Phi(\mathbf{x})$ is a $L \times S$ matrix, having $\Phi_{\ell s} = \phi_s(x_\ell)$. The manifold aligned probabilistic model is a flat mixture model

$$p(\mathbf{t} | \mathbf{W}, \Sigma_\ell) = \frac{1}{L} \sum_{\ell=1}^L p(\mathbf{t} | x_\ell, \Sigma_\ell, \mathbf{W}), \quad (41)$$

where the mixture components are locally manifold-aligned multivariate Gaussians centered at the embedded points $\tilde{\mathbf{t}}_\ell \in \bar{\mathcal{P}}$:

$$p(\mathbf{t} | x_\ell, \Sigma_\ell, \mathbf{W}) = \frac{1}{[(2\pi)^D |\Sigma_\ell|]^{\frac{1}{2}}} \exp \left(-\frac{\Delta \mathbf{t}^\top \Sigma_\ell^{-1} \Delta \mathbf{t}}{2} \right) \quad (42)$$

with $\Delta \mathbf{t} = \mathbf{y}(x_\ell; \mathbf{W}) - \mathbf{t}$. As proposed in Bishop et al. (1998a), we model the local manifold-aligned covariance matrix by computing the derivatives of the mapping function with respect to the latent variables:

$$\Sigma_\ell = \frac{1}{\nu} \mathbf{I} + \omega \frac{\partial \mathbf{y}^\top}{\partial x} \bigg|_{x_\ell} \frac{\partial \mathbf{y}}{\partial x} \bigg|_{x_\ell}, \quad (43)$$

where the purpose of parameter ν (in this work $\nu = 1e3$) is avoiding singularity of Σ_ℓ and ω is a scaling factor equal to the distance between neighboring nodes in the latent space. The derivatives of the mapping function with respect to the latent space coordinates are:

$$\mathbf{g}(x_\ell) = \frac{\partial \mathbf{y}}{\partial x} \bigg|_{x_\ell} = \frac{\partial \Phi}{\partial x} \bigg|_{x_\ell} \mathbf{W} = \sum_{s=1}^S \frac{(x_\ell - x_s)}{\sigma^2} \phi(x_\ell) \mathbf{w}_s \in \mathbb{R}^D, \quad (44)$$

where we denote by \mathbf{w}_s the s -th row of matrix \mathbf{W} . As a latent variable model, Stream GTM can be trained to maximize log-likelihood

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^N \ln \left\{ \frac{1}{L} \sum_{\ell=1}^L p(\mathbf{t}_n | x_\ell, \Sigma_\ell, \mathbf{W}) \right\} \quad (45)$$

via the E-M algorithm outlined in Bishop et al. (1998a). All parameters in SGTm are listed in Table 5. The application of SGTm to the individual manifolds in the synthetic data set results in two probabilistic model which can be visualized by their iso-surfaces corresponding to an iso-value of their Probability Density Functions (PDFs), see Fig. 8, pink surfaces.

From the discussion presented, the toolbox is mainly dependent on one single hyper-parameter: the neighborhood radius r . It is advisable to choose the parameter after visual inspection of the data set at hand. The choice of this hyper-parameter influences the computational cost of the whole methodology, since LAAT, EM3 A, Dimensionality Index, Crawling and SGTm all rely on this for the computation of local PCA. If the radius r is chosen so that a sphere of radius r encloses the estimated thickness of

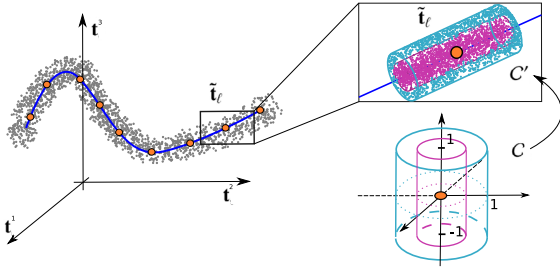


Fig. 9. Sketch depicting the formation of the uniformly sampled concentric cylindrical volumes, aligned with the local tangent space of manifold \mathcal{M}_k on the SGTM center $\tilde{\mathbf{t}}_\ell$.

the filaments within the data set, slight variations of this hyper-parameter from the designated value do not influence the results significantly. A more detailed analysis of the stability of the toolbox w.r.t. r will be presented in an additional work, in preparation. It is not straightforward to estimate the computational cost of the toolbox once the radius r has been chosen, because of the recursive nature of most algorithms. A more detailed analysis of this on specific data sets can be found in the respective papers: (Taghribi et al., 2022) for LAAT, (Mohammadi and Bunte, 2020) and for EM3 A. A detailed theoretical analysis of EM3 A is also presented in Mohammadi et al. (2021) where convergence bounds are outlined together with optimal estimation of hyper-parameters. Quantitative analysis of the effect of changes in the parameters, as well as computational costs for rest of the methodologies (Dimensionality Index, Crawling and SGTM) can be found in Canducci et al. (2022). In the respective papers, each methodology is proofed against state-of-the-art comparable techniques.

4. Visualization techniques

4.1. Bi-dimensional profiles

After optimization of SGTM through the E–M algorithm, every manifold \mathcal{M}_k is represented as a Gaussian mixture with manifold aligned noise, whose updated centers $\{\tilde{\mathbf{t}}_\ell; \ell = 1, \dots, L^k\}$ are constrained to lie on a one-dimensional subspace of \mathbb{R}^3 . We now describe a methodology that, taking full advantage of the probabilistic nature of SGTM, simultaneously recovers the behavior of properties along the manifold's elongation and its thickness within the simulated volume. This methodology gives a comprehensive view of the extracted manifold in a single frame, allowing for a better understanding of its main radial and longitudinal features.

The centers obtained at the end of optimization have most likely shifted along the manifold, due to local variations of point density within the manifold itself. In order to take this shift into consideration we update the tangent bundle to manifold \mathcal{M}_k by computing the derivative of the (trained) mapping function with respect to latent center x_ℓ :

$$\hat{\xi}_\ell = \mathbf{g}(x_\ell), \quad (46)$$

where $\mathbf{g}(x_\ell)$ is derived in Eq. (44). Doing this for every center $x_\ell \in \mathcal{P}^k$ we obtain the updated subset of the tangent bundle $TM^k = \{\hat{\xi}_\ell; \ell = 1, \dots, L^k - 1\}$ of manifold \mathcal{M}_k . Let us now consider a point cloud \mathcal{C} containing points as row vectors, uniformly sampling a cylindrical volume of radius 2, aligned along the z -axis and centered at the origin $\mathcal{O} = (0, 0, 0)$. In order to radially partition the point cloud \mathcal{C} into concentric cylindrical

shells (bottom right panel of Fig. 9, cyan and magenta cylinders), we first create $\bar{\mathcal{C}}$ by projecting \mathcal{C} onto the x - y plane:

$$\bar{\mathcal{C}} = \mathcal{C} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad (47)$$

For every point $\bar{\mathbf{p}} \in \bar{\mathcal{C}}$ we compute its distance to the projected origin $d(\bar{\mathbf{p}}, \bar{\mathbf{0}}) = \|\bar{\mathbf{p}} - [0, 0]\|$. We can now group points in $\bar{\mathcal{C}}$ so that

$$\mathcal{I}_i = \{j \mid d_{i-1}^r \leq d(\bar{\mathbf{p}}_j, \bar{\mathbf{0}}) < d_i^r\} \quad \forall \bar{\mathbf{p}}_j \in \bar{\mathcal{C}}, \quad (48)$$

where $d_i^r = (i-1) \times r_M / (c-1)$ is the low extreme of the interval defined by two consecutive concentric rings on the cylinder, r_M the maximum allowed distance from the mean curve of the manifold, c the desired number of bins across the radial direction and $i = 1, \dots, c-1$. The sets $\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_{N^r}$ contain all indices of points in $\bar{\mathcal{C}}$ (and thus in \mathcal{C}) belonging to specific cylindrical shells concentric with respect to the origin $\bar{\mathbf{0}}$ ($\mathbf{0}$) and the z -axis. It is always possible to scale, translate and rotate the point cloud so that the cylindrical axis is oriented as vector $\hat{\xi}_\ell$, the origin overposed to center $\tilde{\mathbf{t}}_\ell$ and the axis length equal to $d_{\ell, \ell+1}$ (as shown in Fig. 9, top, right panel).

Scaling. In matrix notation, the scaling operator is \mathbf{S}

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d_{\ell, \ell+1} \end{pmatrix}$$

Rotation. We can compute the quaternion \mathbf{q} (Hamilton, 1866) where the first three components are given by $\hat{\xi}_\ell \times (0, 0, 1)$ and the 4-th component by $\hat{\xi}_\ell \cdot (0, 0, 1)$. Its matrix representation is given by \mathbf{R} :

$$\mathbf{R} = \begin{pmatrix} 1 - 2q_2^2 - 2q_3^2 & 2q_1q_2 - 2q_3q_4 & 2q_1q_3 + 2q_2q_4 \\ 2q_1q_2 + 2q_3q_4 & 1 - 2q_1^2 - 2q_3^2 & 2q_2q_3 - 2q_1q_4 \\ 2q_1q_3 - 2q_2q_4 & 2q_2q_3 + 2q_1q_4 & 1 - 2q_1^2 - 2q_2^2 \end{pmatrix}$$

Shift. We can then shift the scaled and rotated point cloud so that its origin is on center $\tilde{\mathbf{t}}_\ell$.

Any point $\mathbf{p} \in \mathcal{C}$ is then mapped to $\mathbf{p}' \in \mathcal{C}'$ under the combined operator as:

$$\mathbf{p}' = (\mathbf{p} \mathbf{S}) \mathbf{R} + \tilde{\mathbf{t}}_\ell. \quad (49)$$

Having obtained a point cloud \mathcal{C}' uniformly sampling a thick cylindrical volume with axis tangential to manifold \mathcal{M}_k on point $\tilde{\mathbf{t}}_\ell$, we can now compute the weighted mean of any quantity contained in the data set, over the volume sampled by \mathcal{C}' .

Note that the families of indices $\mathcal{I}_1, \dots, \mathcal{I}_{N^r}$, when applied to \mathcal{C}' , contain all indices of points in cylindrical shells concentric with respect to point $\tilde{\mathbf{t}}_\ell$ and vector $\hat{\xi}_\ell$, radially partitioning the cylindrical volume sampled by \mathcal{C}' . For each linear segment of manifold \mathcal{M}_k , parameterized by its corresponding SGTM, we have now obtained a uniformly sampled cylindrical volume aligned along its corresponding local tangent space. Computing $\langle T_m(\mathbf{p}') \rangle$ for every $\mathbf{p}' \in \mathcal{C}'$ we can now evaluate the mean value of T_m over the concentric rings defined by the index families $\mathcal{I}_1, \dots, \mathcal{I}_{N^r}$ as:

$$\overline{T_{m,i}} = \langle T_m(d_{i-1}^r, d_i^r) \rangle = \frac{\sum_{j \in \mathcal{I}_i} \langle T_m(\mathbf{p}'_j) \rangle}{|\mathcal{I}_i|}, \quad (50)$$

obtaining the mean of T_m over the cylindrical shell between (d_{i-1}^r, d_i^r) for every $i = 1, \dots, N^r$.

We can iterate the whole process for every center of SGTM, obtaining for each linear segment, the distribution of T_m in concentric cylindrical shells centered on the current center (Fig. 9, left panel). By considering both longitudinal (defined recursively by the centers of SGTM) and radial (obtained by the linear operator defined in Eq. (49) on point cloud \mathcal{C}) profiles, we obtain the

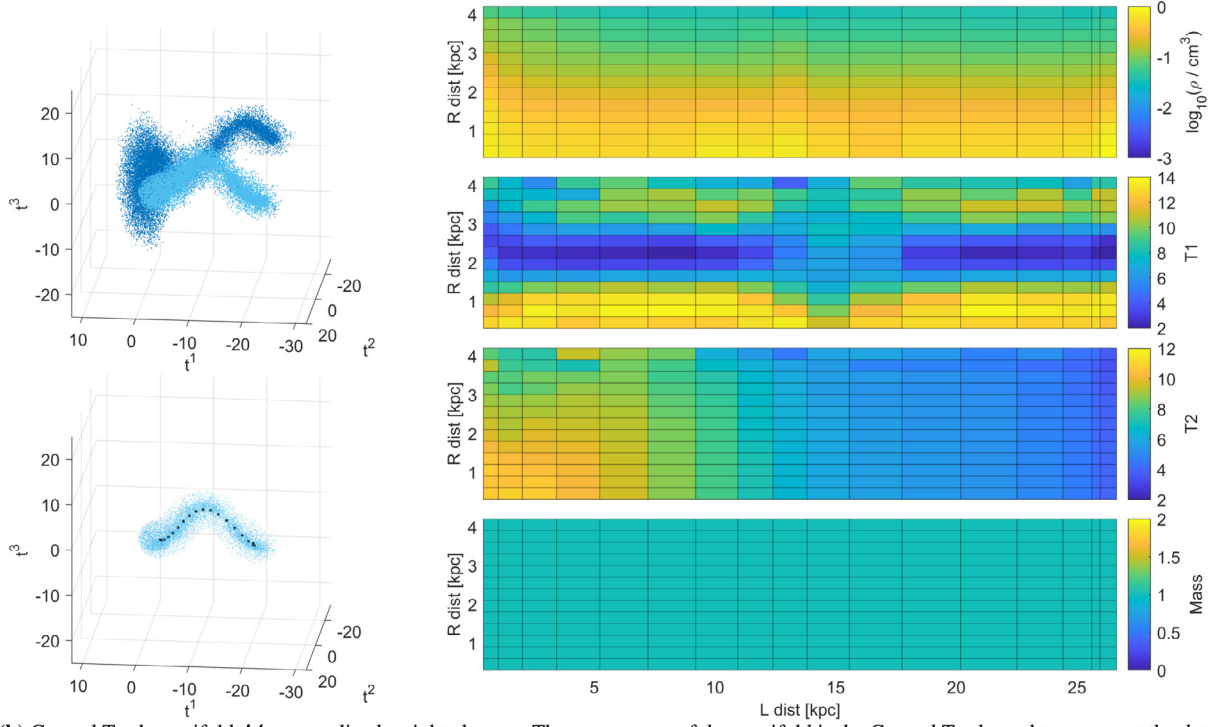
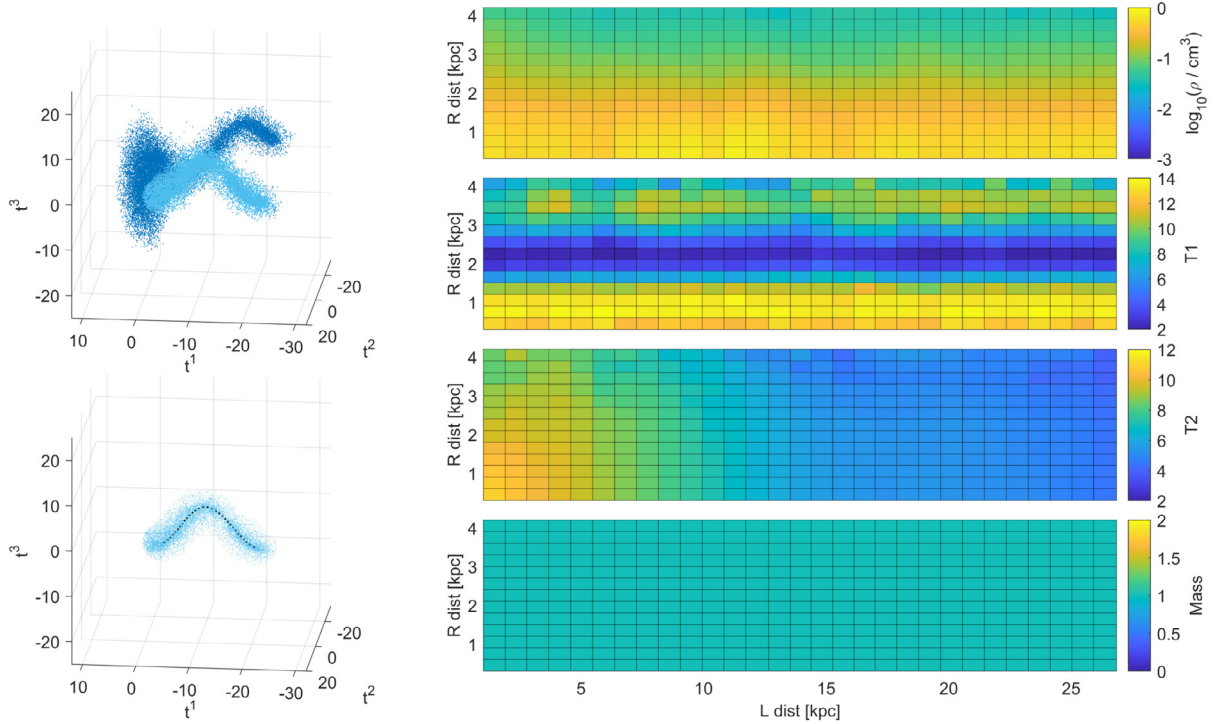
(a) Estimated manifold \mathcal{M}_3 , normalized weighted mean, as recovered after the application of the toolbox on the initial noisy data set.(b) Ground Truth manifold \mathcal{M}_3 , normalized weighted mean. The mean curve of the manifold is the Ground Truth used to construct the data set.

Fig. 10. Comparison of the estimated manifold \mathcal{M}_3 (a) and its Ground Truth (b) on the synthetic Jellyfish data. Left: top panels show the noisy data set points (blue) with background noise removed and including the noisy manifold identification (cyan) both for SGTM (a) and the Ground Truth (b). Correspondingly, the bottom panels depict the skeleton of \mathcal{M}_3 (black) recovered via Crawling and SGTM in (a) and its Ground Truth in (b). Right: contains the estimated (a) and true (b) bi-dimensional profiles (normalized with Eq. (60)) for variables $\bar{\rho}_i$, \bar{T}_{1i} , \bar{T}_{2i} , and \bar{m}_i (from top to bottom).

plots shown in Figs. 10a and 11a. In both panels, the vertical axis of each plot contains the radius of the cylindrical shells d^r and the horizontal axis the approximated geodesic distance (computed by summation of the lengths of the individual linear segments) from the head of the manifold. From these plots we can verify that the behavior of quantities T_1 (second panel from top) and T_2 (third

panel from top) are in agreement with how they were designed when constructing the data set (Section 2). The decreasing profile for quantity T_1 and increasing for T_2 along manifold \mathcal{M}_2 is detected, as well as the sinusoidal behavior of quantity T_1 along manifold \mathcal{M}_3 's thickness and T_2 's decreasing profile along its longitudinal elongation. The bottom plot in each panel presents

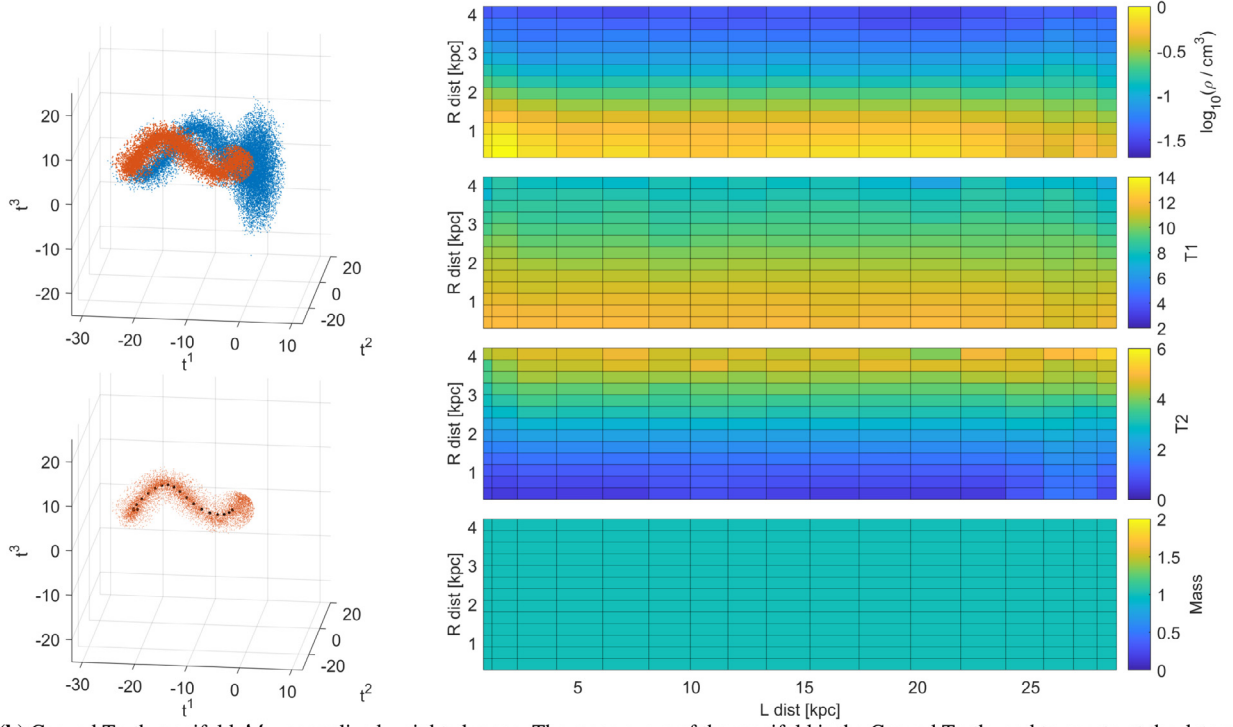
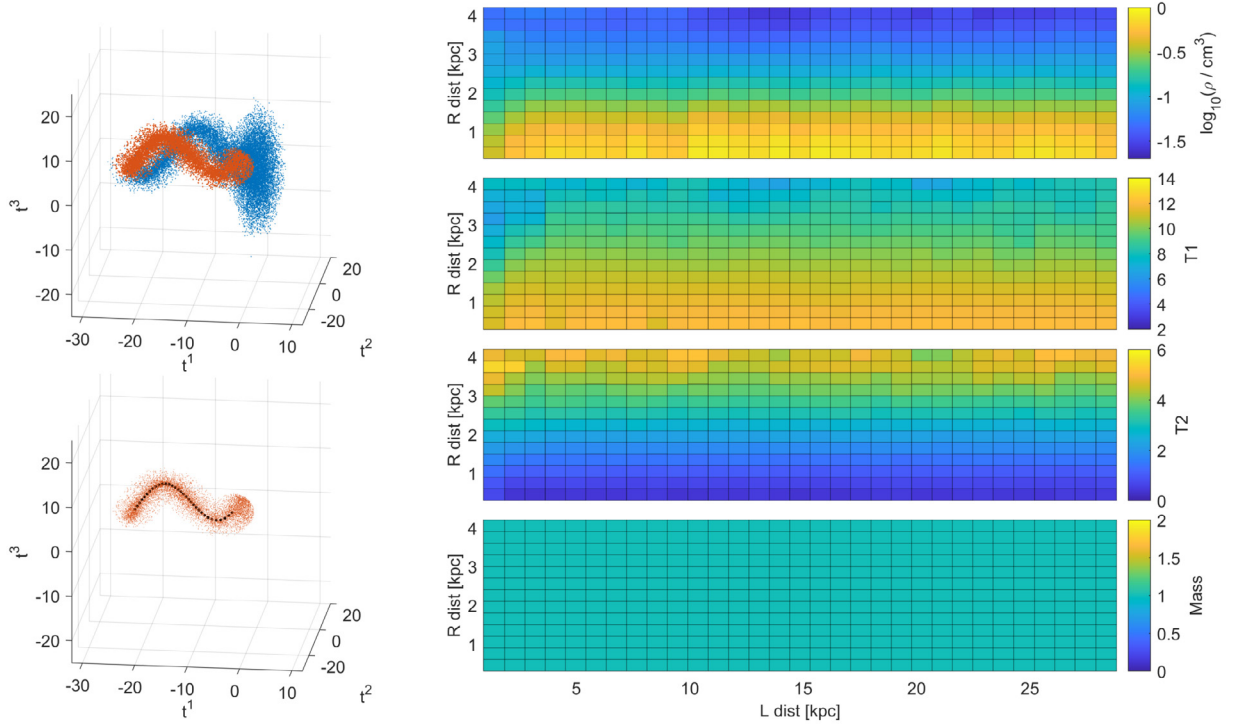
(a) Estimated manifold \mathcal{M}_2 , normalized weighted mean, as recovered after the application of the toolbox on the initial noisy data set.(b) Ground Truth manifold \mathcal{M}_2 , normalized weighted mean. The mean curve of the manifold is the Ground Truth used to construct the data set.

Fig. 11. Comparison of the estimated manifold \mathcal{M}_2 (a) and its Ground Truth (b) on the synthetic Jellyfish data. Left: top panels show the noisy data set points (blue) with background noise removed and including the noisy manifold identification (orange) both for SGTM (a) and the Ground Truth (b). Correspondingly, the bottom panels depict the skeleton of \mathcal{M}_2 (black) recovered via Crawling and SGTM in (a) and its Ground Truth in (b). Right: contains the estimated (a) and true (b) bi-dimensional profiles (normalized with Eq. (60)) for variables $\bar{\rho}_i$, \bar{T}_{1i} , \bar{T}_{2i} , and \bar{m}_i (from top to bottom).

the mass distribution over the radial and longitudinal dimensions of the manifolds. As expected, the mass is constant throughout the sampled volumes and it is everywhere $\bar{m}_i = 1$.

For each manifold, Figs. 10b and 11b show the true profiles recovered by using the ground truth skeletons described in Section 2. As previously described, variable T_1 shows a sinusoidal

variation along the radial direction of manifold \mathcal{M}_3 and decreasing radial profile across manifold \mathcal{M}_2 , variable T_2 is decreasing on the longitudinal direction of manifold \mathcal{M}_3 and radially decreasing from the core of manifold \mathcal{M}_2 . The profiles, when compared with the ones obtained using the skeletons recovered by our

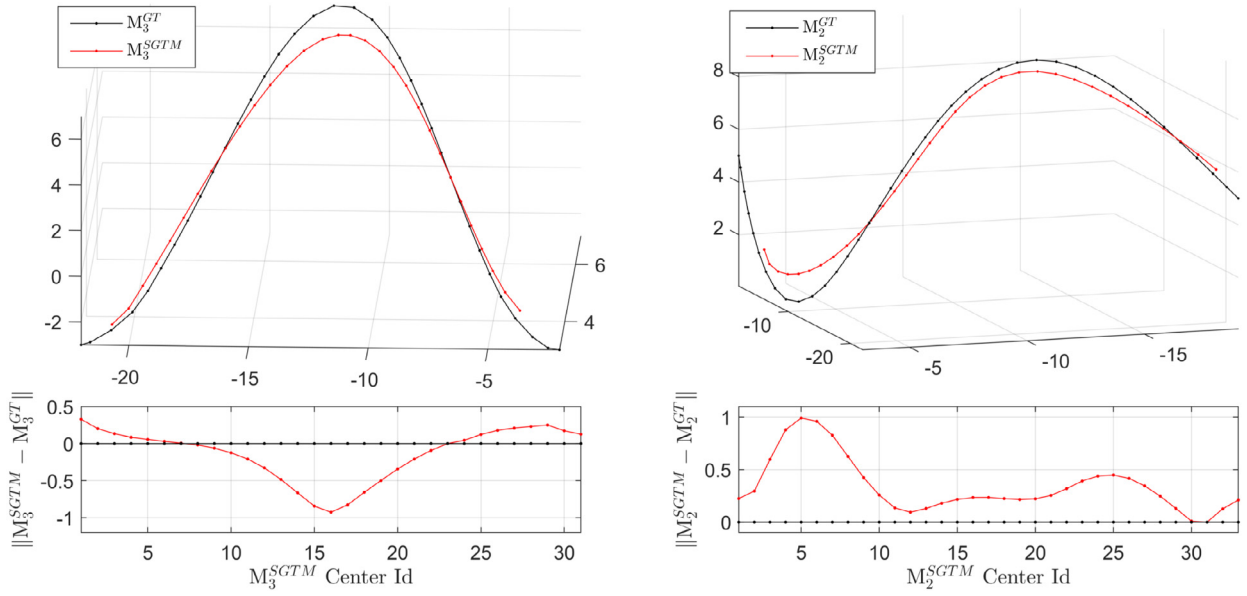


Fig. 12. Top row: recovered manifolds (in red) via SGTM and their corresponding Ground Truth (black). Bottom row: Point-by-point orthogonal distance between \mathcal{M}_k^{GT} and \mathcal{M}_k^{SGTM} .

methodology, look virtually identical. Small deviations are noticeable for variable T_1 halfway through manifold \mathcal{M}_3 , however even in this case the variation is minimal and does not compromise the overall agreement. This demonstration is a first quantitative confirmation of the accuracy of our methodology in recovering the underlying structures of the data set. Having obtained a parameterization of the two manifolds in the data set (\mathcal{M}_k^{SGTM}), it is now possible to compare the recovered structures with their Ground Truth (\mathcal{M}_k^{GT} , given in Eq. (1)). In order to have a measure that is unbiased with respect to the particular parameterization produced by Crawling and SGTM, we re-sample each manifold by projecting the Ground Truth points onto the recovered curves, along their local orthogonal planes. The module of the projection gives the local orthogonal distance between a point in \mathcal{M}_k^{GT} and its corresponding in \mathcal{M}_k^{SGTM} . The two recovered manifolds (red points), together with their Ground Truths (black points) are shown in the top row of Fig. 12. The point-by-point orthogonal distance is shown in the bottom row. The maximum distance between the two curves is at the point of maximal curvature of the manifold. This discrepancy has two main contributions. Firstly, the data set obtained with EM3 A is slightly misaligned with the Ground Truth. This is probably due to the random sparsity of the transverse and background noise in the data set. The combined effect of these two uncertainty factors may affect the local density estimation, and thus the displacement of nearby points onto local tangent spaces (over- or under-shooting). The second contribution is due to the application of Crawling and SGTM. While the initialization provided by crawling lies on the data set obtained with EM3 A and carries the same uncertainty, the SGTM formulation should compensate for possible small deviations. However, the radius $r = 2$ chosen in this example might be too large for capturing the high-curvature regions. Also, the imposed σ value (estimated from the RBF setup) used for constraining the model's complexity is possibly too large. It is however generally advisable to avoid over-fitting the data with an overly complex model for the sake of its generalizability to unseen data sampled from the same distribution. Despite this minor discrepancy between the recovered curves, an overall agreement is achieved for both manifolds. The discrepancy between the two curves at the peak of the distance is also responsible for the deformation of the bi-dimensional profiles, especially in the central part of the plots

Table 6

Recovery accuracy from bi-dimensional profiles.

Manifold	$\zeta^{0.2}(\log^{10} \rho)$	$\zeta^{0.2}(T_1)$	$\zeta^{0.2}(T_2)$
\mathcal{M}_3	0.9897	0.8615	0.9846
\mathcal{M}_2	0.9089	1.0000	0.8776

(e.g. Fig. 10 panels a and b) for variable T_1 (sinusoidal profile from the core of manifold \mathcal{M}_2). Nonetheless, despite the amount of noise corrupting both the structures and the simulated quantities, 1-DREAM still manages to recover a reasonable approximation to the Ground Truth.

The centers obtained by projection of the Ground-Truth onto the manifolds recovered via SGTM can also be used to obtain new Ground Truth bi-dimensional profiles. In this case, the two bi-dimensional profiles for each manifold are directly comparable. The mean value of variable T_m in bin (d_{i-1}^r, d_i^r) of the bi-dimensional profile $\bar{T}_{m,i} = \langle T_m(d_{i-1}^r, d_i^r) \rangle$ is derived according to Eq. (50) for the ground-truth, $\bar{T}_{m,i}^{GT}$, and SGTM, $\bar{T}_{m,i}^{SGTM}$, recovered manifolds. For each bin we can now compute the fractional deviation of variable T_m between the two recovered structures:

$$\zeta_{m,i} = \left\| \frac{\bar{T}_{m,i}^{SGTM} - \bar{T}_{m,i}^{GT}}{\bar{T}_{m,i}^{GT}} \right\|. \quad (51)$$

We estimate the accuracy of the recovered bi-dimensional profiles as the ratio of pixels having fractional deviation lower than 0.2 and the total number of pixels, such that

$$\zeta^{0.2}(T_m) = \frac{| \{ (i, \ell) \mid \zeta_{m,i} < 0.2 \} |}{N^r \times (L^k - 1)}. \quad (52)$$

The values of $\zeta^{0.2}(T_m)$, T_m being $\log_{10}(\rho)$, T_1 or T_2 for both manifolds \mathcal{M}_2 and \mathcal{M}_3 are given in Table 6. In agreement with the previous discussion, variable T_1 for manifold \mathcal{M}_3 and T_2 for \mathcal{M}_2 show the largest variation with respect to the Ground-Truth, however, only $\sim 14\%$ and $\sim 13\%$ (respectively) of the pixels have larger fractional deviation than 0.2, while we reach a good agreement between Ground-Truth and SGTM in the other cases. The parameters of the Bi-dimensional profile technique are provided in Table 7.

Table 7

Full list of parameters for Bi-dimensional profiles.

$c^* \in \mathbb{N}$	Radial number of bins
$r_M^* \in \mathbb{R}$	Maximum radial distance from mean curve

4.2. Co-moving orthonormal coordinate frames

For each manifold \mathcal{M}_k , we can now obtain a better discretization of its “spine”, by introducing more points in the latent space and propagating them in the ambient space through the mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W}) : \mathcal{P}^k \rightarrow \overline{\mathcal{P}}^k$. Consider the linear segment $\mathcal{I}_\ell := [x_\ell, x_{\ell+1}]$, where $x_\ell, x_{\ell+1} \in \mathcal{P}^k$. Denoting by $d_\ell = \|\mathcal{I}_\ell\|$ the length of segment \mathcal{I}_ℓ , let us assume that the number of equidistant points to be inserted in \mathcal{I}_ℓ is N_ℓ . We can define the new points via the recursive rule:

$$x_\ell^m = x_\ell + \frac{m}{N_\ell + 1} d_\ell \quad \text{for } m = 1, \dots, N_\ell. \quad (53)$$

Applying this relation to points $x_\ell, \forall \ell = 1, \dots, L^k - 1$, We obtain the up-sampled latent space $\mathcal{P}_\uparrow^k = \{x_1, x_1^1, \dots, x_1^{N_1}, x_2, \dots, x_{L^k-1}, x_{L^k-1}^1, \dots, x_{L^k-1}^{N_{L^k-1}}, x_{L^k}\}$, having size $|\mathcal{P}_\uparrow^k| = N^\ell (L^k - 1) + L^k$. Propagation of latent point set \mathcal{P}_\uparrow^k into the ambient space through mapping function $\mathbf{y}(x_\ell; \mathbf{W})$, for every $x_\ell \in \mathcal{P}_\uparrow^k$, leads to the up-sampled embedded point-set $\overline{\mathcal{P}}_\uparrow^k$, as depicted in Fig. 13, right panel, black dots. As in Section 4.1, the tangent bundle TM^k is updated by applying Eq. (46) to every point in $\overline{\mathcal{P}}_\uparrow^k$. In this section, slightly abusing mathematical notation, we will use index ℓ for any latent (and corresponding embedded) point belonging to \mathcal{P}_ℓ^k and drop subscript \uparrow , for readability purposes.

For every $\hat{\xi}_\ell \in TM^k$ we can recover a set of two vectors, $\mathbf{u}_1 = (u_1^1, u_1^2, u_1^3)$ and $\mathbf{u}_2 = (u_2^1, u_2^2, u_2^3)$, perpendicular to $\hat{\xi}_\ell$, spanning the perpendicular plane \mathcal{T}_ℓ^\perp to manifold \mathcal{M}_k on center $\tilde{\mathbf{t}}_\ell$. This is achieved by solving the system of linear equations given by:

$$\begin{cases} \mathbf{u}_1 \cdot \mathbf{u}_2 = u_1^1 u_2^1 + u_1^2 u_2^2 + u_1^3 u_2^3 = 0 \\ \mathbf{u}_1 \cdot \hat{\xi}_\ell = u_1^1 \hat{\xi}_\ell^1 + u_1^2 \hat{\xi}_\ell^2 + u_1^3 \hat{\xi}_\ell^3 = 0 \\ \mathbf{u}_2 \cdot \hat{\xi}_\ell = u_2^1 \hat{\xi}_\ell^1 + u_2^2 \hat{\xi}_\ell^2 + u_2^3 \hat{\xi}_\ell^3 = 0 \end{cases} \quad (54)$$

Being a degenerate system of linear equations we can recover an infinite number of solutions giving infinite pairs of vectors spanning the perpendicular plane \mathcal{T}_ℓ^\perp . In order to maintain consistency throughout the manifold's elongation, we choose the solution to be:

$$\mathbf{u}_1 = (\hat{\xi}_\ell^2, -\hat{\xi}_\ell^1, 0), \quad (55)$$

$$\mathbf{u}_2 = (\hat{\xi}_\ell^1 \hat{\xi}_\ell^3, \hat{\xi}_\ell^2 \hat{\xi}_\ell^3, -[(\hat{\xi}_\ell^1)^2 + (\hat{\xi}_\ell^2)^2]), \quad (56)$$

so that $\mathcal{T}_\ell^\perp = \text{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2)$, where $\hat{\mathbf{u}}_1 = \mathbf{u}_1 / \|\mathbf{u}_1\|$ and $\hat{\mathbf{u}}_2 = \mathbf{u}_2 / \|\mathbf{u}_2\|$. Under this scheme, the two vectors form an orthonormal coordinate frame for the plane locally perpendicular to center $\tilde{\mathbf{t}}_\ell$. The two vectors are shown in Fig. 13 as the magenta and blue arrows, changing direction slightly, between any pair of adjacent centers in \mathcal{P}^k . The tangent bundle is here also shown (sampled on points in \mathcal{P}^k) as green arrows. We can now impose a regular $M \times M$ square grid of side a on plane \mathcal{T}_ℓ^\perp , taking advantage of the local coordinate frame given by unit vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ (black gridded rotated squares in Fig. 13). We define the set $\mathcal{Y}_\ell = \{\mathbf{y}_{11}, \dots, \mathbf{y}_{1M}, \mathbf{y}_{21}, \dots, \mathbf{y}_{2M}, \dots, \mathbf{y}_{M1}, \dots, \mathbf{y}_{MM}\}$, where

$$\begin{cases} \mathbf{y}_{ij} = \mathbf{0}_\ell + i \delta \hat{\mathbf{u}}_1 + j \delta \hat{\mathbf{u}}_2; \\ \mathbf{0}_\ell = \tilde{\mathbf{t}}_\ell - 2(\hat{\mathbf{u}}_1 + \hat{\mathbf{u}}_2) \end{cases} \quad (57)$$

and the increments along vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ are given by $\delta \hat{\mathbf{u}}_1 = (a/M) \hat{\mathbf{u}}_1$ and $\delta \hat{\mathbf{u}}_2 = (a/M) \hat{\mathbf{u}}_2$ respectively. The number of bins

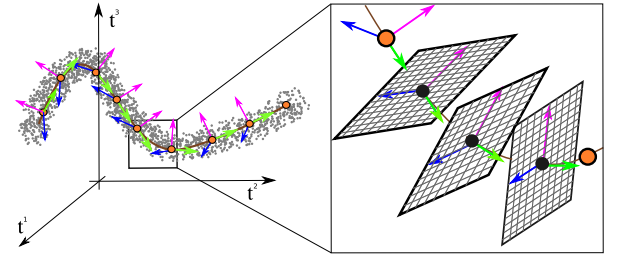


Fig. 13. Sketch depicting the formation of co-moving orthonormal planes along a portion of manifold \mathcal{M}_k , enclosed within two centers of the corresponding SGTMs.

M must be chosen as an odd integer in order for $\tilde{\mathbf{t}}_\ell$ to be on the origin of the local coordinate frame. In order to represent only local properties of the manifold we perform a selection of relevant particles in the data set based on their position with respect to the new reference frame. We first compute the projection \mathbf{t}_m^\parallel of all particle's original positions \mathbf{t}_m onto the tangent vector $\hat{\xi}_\ell$ (note that $\|\hat{\xi}_\ell\| = 1$) to manifold \mathcal{M}_k at point $\tilde{\mathbf{t}}_\ell$: $\mathbf{t}_m^\parallel = [(\mathbf{t}_m - \tilde{\mathbf{t}}_\ell) \cdot \hat{\xi}_\ell] \hat{\xi}_\ell$. We assume that the distance \bar{d}_ℓ between adjacent points in $\overline{\mathcal{P}}_\ell^k$ is always proportional to the distance d_ℓ between corresponding adjacent points in \mathcal{P}_ℓ^k and constant.² We then select only those particles such that $\|\mathbf{t}_m^\parallel\| \leq \bar{d}_\ell/2$. Additionally, we compute the perpendicular component \mathbf{t}_m^\perp of position \mathbf{t}_m , by building the projection operator onto \mathcal{T}_ℓ^\perp as in Section 3.2.2: $\mathbf{P} = \mathbf{V}\mathbf{V}^\dagger$, where \mathbf{V} is the matrix having \mathbf{u}_1 and \mathbf{u}_2 as column vectors: $\mathbf{t}_m^\perp = \mathbf{P}(\mathbf{t}_m - \tilde{\mathbf{t}}_\ell)$. In our analysis we will only consider particles lying within the sphere of radius b , centered on $\tilde{\mathbf{t}}_\ell$: $\|\mathbf{t}_m^\perp\| \leq b$, where b can be imposed by the user or automatically selected as $b = a\sqrt{2}/2$: the half-diagonal of the gridded plane \mathcal{T}_ℓ^\perp . However, this parameter is only defined when a non-SPH weighting (e.g. Gaussian) scheme is applied. In fact, when SPH is in place, there is no need to select a subset of particles surrounding the plane in order to compute the mean value of properties on the plane. This is achieved via the SPH weighting scheme presented in Eq. (60) through the smoothing length parameter, defined for all particles in the data set. Identifying with \mathcal{J} the index set of all particles satisfying these two conditions, we can now compute the weighted mean of variables $T_1(\mathcal{J}, \mathbf{y}_{ij})$ and $T_2(\mathcal{J}, \mathbf{y}_{ij})$ with respect to points $\mathbf{y}_{ij} \in \mathcal{Y}_\ell$, under the SPH formulation (see Section 4.3). Top row of Figs. 14a–14b present the behaviors of variables T_1 , T_2 and density ρ respectively, for manifold \mathcal{M}_2 detected in the synthetic data set, computed at position $\tilde{\mathbf{t}}_\ell^k$ ($\tilde{\mathbf{t}}_\ell^k$ varying along the manifolds for each group of pictures). Having the probabilistic model for manifold \mathcal{M}_k as SGTMs, we also compute the Probability Density Function of the Mixture on plane \mathcal{T}_ℓ^\perp as:

$$p^k(\mathbf{y}_{ij}) = \sum_{x_\ell \in \mathcal{P}^k} \pi_\ell p(\mathbf{y}_{ij} | x_\ell, \hat{\Sigma}_\ell, \hat{\mathbf{W}}), \quad (58)$$

where $\hat{\Sigma}_\ell$ and $\hat{\mathbf{W}}$ are the parameters found through optimization with the EM algorithm (see Section 3.2.3) and π_ℓ the component proportion given by $\pi_\ell = 1/|\mathcal{P}^k|$.

The PDFs for manifolds \mathcal{M}_k on plane \mathcal{T}_ℓ^\perp at position \mathbf{t}_ℓ^k are shown in the bottom left panel of Figs. 14a–14b, while central and

² This might not always be the case, but present small variations due to training SGTMs. However, since we interpolate the SGTMs model by up-sampling \mathcal{P}_ℓ^k in order to increase smoothness of the orthonormal planes, the distance between adjacent centers in \mathcal{P}_ℓ^k (and thus its embedding $\overline{\mathcal{P}}_\ell^k$) can always be regularized by up-sampling this set more densely.

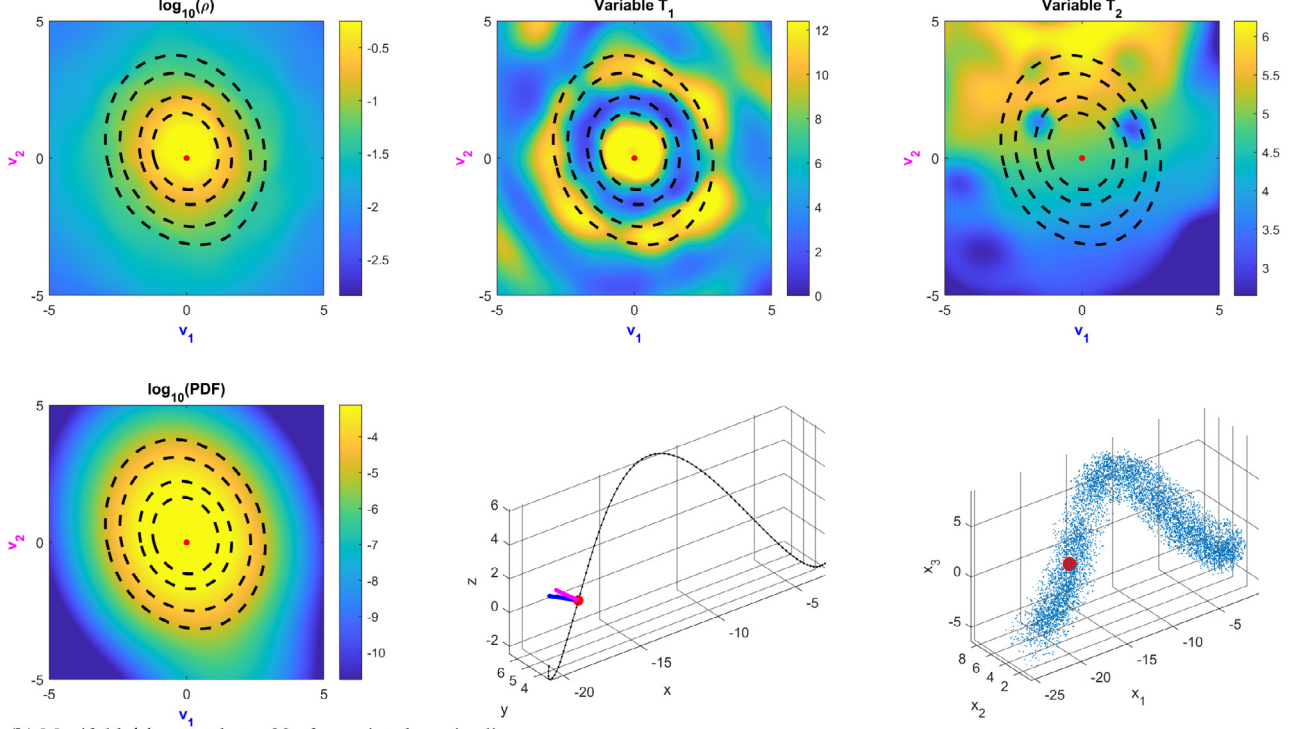
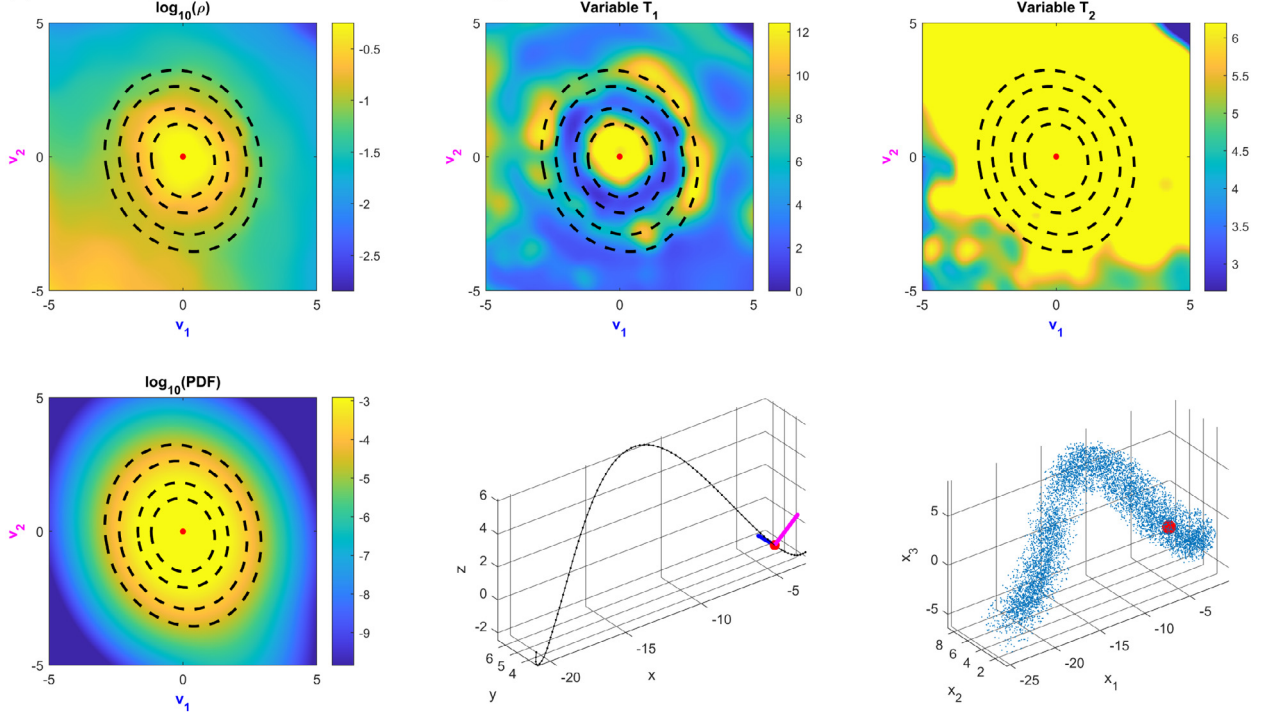
(a) Manifold \mathcal{M}_3 , snapshot n.20 of associated movie clip.(b) Manifold \mathcal{M}_3 , snapshot n.80 of associated movie clip.

Fig. 14. Snapshots n.20 (a) and n.80 (b) of the video clip generated by considering co-moving coordinate frames on the elongation of manifold \mathcal{M}_3 . The top row panels show the distribution of variables ρ , T_1 and T_2 over the coordinate frame \mathcal{T}_ℓ^\perp defined on \mathbf{t}_ℓ , shown in bottom central and right panels as a red sphere. Bottom left panel presents the Probability Density Function obtained through SGTM on the same plane.

right panels show the current position on the manifold and with respect to the whole data set respectively. By applying recursively this procedure to every $\mathbf{t}_\ell \in \overline{\mathcal{P}}^k$ we obtain a representation of the variables of interest, on co-moving orthonormal reference frames along each manifold, showing the clear radial behavior of these variables.

Figs. 14a–14b present the results for selected points on manifold \mathcal{M}_3 . The top central panel in each figure clearly manifests

the sinusoidal radial behavior of variable T_1 . This implies that the center of the manifold has been correctly identified by SGTM and its centers lie closely to its underlying nature (see Eq. (1)). Comparing top right panel in Fig. 14a and the one in Fig. 14b, the longitudinal dimming of variable T_2 can be verified, showing that its true nature has been correctly recovered. The snapshots obtained via the co-moving orthonormal coordinate frames technique are joined sequentially into a movie that shows the

Table 8

Full list of parameters for Co-moving Orthonormal coordinate frames.

$N_\ell \in \mathbb{N}$ ($N_\ell = 5$)	Latent interval upsample size
$M \in \mathbb{N}$ ($M = 25$)	Pixels on plane
$a^* \in \mathbb{R}$ ($a > 0$)	Length of plane
$b \in \mathbb{R}$ ($b = a\sqrt{2}/2$)	Distance from plane

evolution of the quantities while moving along the manifold. The presented figures only show selected snapshots of the movie for manifold \mathcal{M}_3 , however the analysis has been performed for both manifolds in the synthetic data sets and the associated movies can be found at: <https://git.lwp.rug.nl/cs.projects/1DREAM>. The snapshots presented in this work are individual frames of the movie clip and they are referred to as “snaphsot n. ...” when addressed in the captions (see e.g. Table 8). A list of parameters used in this case for the described visualization technique is given in Table 8.

4.3. Weighting schemes

Depending on the data at hand, it is possible to implement multiple different weighting schemes. Here, we focus on two main approaches: an SPH formulation and a classical Gaussian smoothing, with variable length-scale. If the data set is obtained via an SPH simulation, the weighted mean of any simulated quantity has to be computed following the formulation of the code used. In the following, we will consider the weighting scheme implemented in GADGET2 and a general Gaussian smoothing technique. We provide a detailed description of the two formulations for a given center of SGTm and the associated point-set on whose points the mean is intended to be computed (either the uniformly distributed cylindrical shells or the sampled perpendicular plane centered on $\tilde{\mathbf{t}}_\ell$). We will identify the query point-set by $\mathcal{C}' = \{\mathbf{p}'_1, \mathbf{p}'_2, \dots, \mathbf{p}'_N\}$.

4.3.1. SPH-like weights

Consider a point $\mathbf{p}'_i \in \mathcal{C}'$. We need to compute the weighted mean, under the SPH formalism, of variable T_m summing through all the particles $\mathbf{t}_j \in \mathcal{Q}$. The spline approximation of the Gaussian kernel (usually referred to as *smoothing kernel*) on a finite support is:

$$W(q_j, h_j) = \frac{1}{\pi h_j^3} \begin{cases} \frac{1}{4}(2 - q_j)^3 - (1 - q_j)^3 & 0 \leq q_j < 1 \\ \frac{1}{4}(1 - q_j)^3 & 1 \leq q_j < 2 \\ 0 & q_j \geq 2 \end{cases} \quad (59)$$

where $q_j = \|\mathbf{t}_j - \mathbf{p}'_i\|/h_j$ (h_j being the smoothing length defined in Section 2.1). Using the kernels, the exact weighted mean of quantity T_m at point \mathbf{p}' is:

$$\bar{T}_{m,i} := \langle T_m(\mathbf{p}'_i) \rangle = \frac{\sum_{\mathbf{t}_j \in \mathcal{Q}} \frac{m_j}{\rho_j} T_m(\mathbf{t}_j) W(q_j, h_j)}{\sum_{n=1}^{|\mathcal{Q}|} \frac{m_n}{\rho_n} W(q_n, h_n)} \quad (60)$$

The term in the denominator is generally considered to be approximating unity when the particles in a data set are distributed uniformly; however, this is not often the case in practice. Each particle \mathbf{t}_j of an SPH data set samples a spherical volume of radius h_j . Since all particles are evolved following the equations of motion defined by the Lagrangian formulation of fluids, their distribution at a given evolutionary stage is far from uniform. Thus, the *approximation to unity* assumption is generally incorrect at an advanced evolutionary stage. The role of the normalization term $\sum_{n=1}^{|\mathcal{Q}|} \frac{m_n}{\rho_n} W(q_n, h_n)$ in the denominator of Eq. (60) is to eliminate the interpolation's dependence from the particles' distribution.

As such, it cannot be disregarded when computing the weighted mean of quantity T_m on any point in \mathcal{C}' .

In order to check the influence of the normalization term, we show in Fig. 15 panel a and b the corresponding behaviors of $\bar{\rho}_i$, $\bar{T}_{1,i}$, $\bar{T}_{2,i}$ and \bar{m}_i when the normalization term is omitted in Eq. (60). While the ranges of the first three variables exceed their true respective values, it is striking the difference of the mass distribution with respect to the normalized versions (Figs. 10a and 11a). The radial sinusoidal profile of variable T_1 on manifold \mathcal{M}_3 is also completely lost (right side, second panel from the top), as the sparser distribution in the further regions from the core of the manifold does not carry enough weight to compensate for the inner regions.

4.3.2. Gaussian smoothing

If the data set has not been generated via an SPH simulation, the weighted mean of any variable on point-set \mathcal{C}' is obtained by imposing a Gaussian isotropic kernel on each point $\mathbf{p}'_i \in \mathcal{C}'$, with scale length κ , obtaining for each point $\mathbf{t}_j \in \mathcal{Q}$, weight:

$$p(\mathbf{t}_j|\mathbf{p}'_i, \kappa) = \frac{1}{2\pi\kappa^2} \exp\left(-\frac{\|\mathbf{t}_j - \mathbf{p}'_i\|^2}{2\kappa^2}\right). \quad (61)$$

The weighted mean of variable T_m on every point $\mathbf{p}'_i \in \mathcal{C}'$ is then obtained by summing through all particles in data set \mathcal{Q} as:

$$\bar{V}_{m,1} := \langle T_m(\mathbf{p}'_i) \rangle = \frac{\sum_{j=1}^{|\mathcal{Q}|} p(\mathbf{t}_j|\mathbf{p}'_i, \kappa) T_m(\mathbf{t}_j)}{\sum_{n=1}^{|\mathcal{Q}|} p(\mathbf{t}_n|\mathbf{p}'_i, \kappa)}. \quad (62)$$

5. Experiments on different data sets

This section is devoted to the application of the proposed methodology to three different data sets. In order to avoid confusion, each subsection indicates the data set by the same notation used in the previous sections (\mathcal{Q} and $\bar{\mathcal{Q}}$ for the noisy and diffused data sets respectively). The first data set (Section 5.1) is a simulated dwarf galaxy interacting with its host galaxy cluster. In particular, we examine a single simulated snapshot of the dwarf's evolution. The choice of the snapshot is motivated by the presence of multiple gaseous filamentary structures, located mainly at the back of the simulated box and forming a gaseous tail. In this case we are only considering the gas particles' distribution, disregarding Dark Matter and Stellar particles. Here, we analyze the gas temperature (T), gas density (ρ), neutral gas fraction (which is the ratio of neutral, or atomic, gas mass to total gas mass) and metallicity (the iron abundance $[\text{Fe}/\text{H}]$) of the extracted streams. We note that for $[\text{Fe}/\text{H}]$, which is defined as a logarithmic ratio of concentrations, we first recover the linear ratio then we use Eq. (60), and finally we go back to logarithmic scale. The second data set (Section 5.2) is obtained via a Dark Matter simulation of a sample volume of the Universe's Large Scale Structure (LSS). We focus here on the kinematic properties of the filaments of dark matter extracted from the Cosmic Web. The third and last data set is the observed stellar spatial distribution of a sky-region enclosing the globular cluster ω -Centauri. We aim at recovering the two stellar streams detected by Ibata et al. (2019a) and describing their radial and longitudinal density profiles.

5.1. Simulated jellyfish: dwarf galaxy in Fornax cluster

Methodologies: LAAT \rightarrow EM3 A \rightarrow Dimensionality Index \rightarrow Crawling \rightarrow SGTm \rightarrow Bi-dimensional profiles \rightarrow Co-moving Orthonormal coordinate frames.

The simulations initially consider a dwarf galaxy evolving in isolation for 8 billion years. Here, isolation means that the galaxy

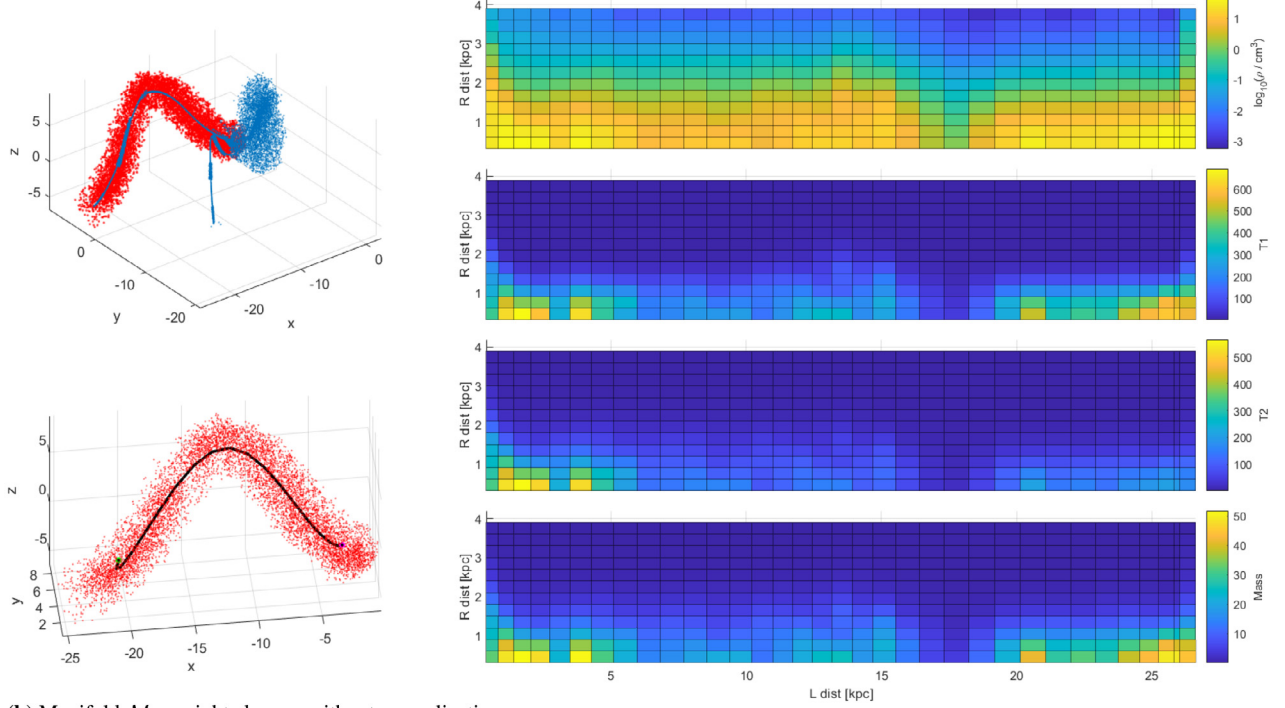
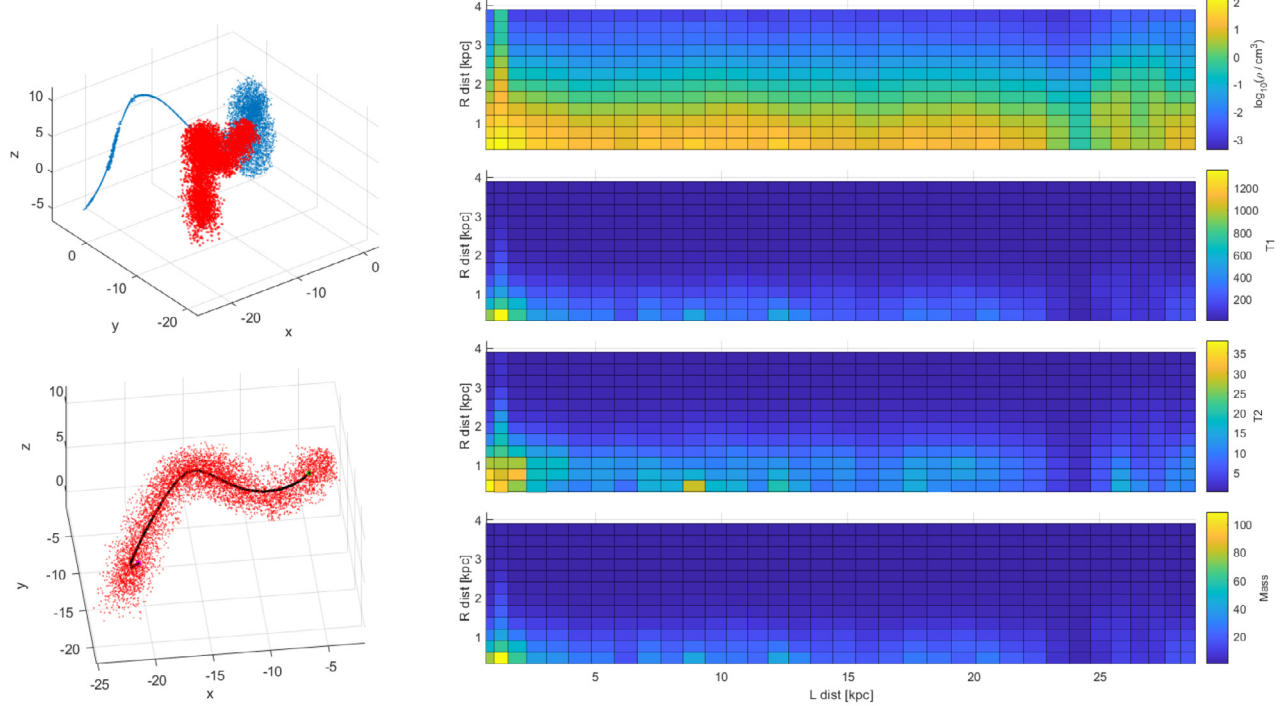
(a) Manifold M_3 , weighted mean without normalization.(b) Manifold M_2 , weighted mean without normalization.

Fig. 15. Same as panels Figs. 10a and 11a but without normalization in the SPH weighting scheme.

was assembling its mass through mergers but was not absorbed by a more massive structure, such as a galaxy cluster, where its internal properties could be affected by external processes such as gravitational interactions with other galaxies and ram-pressure stripping. A full catalogue and detailed study of these galaxies can be found in Verbeke et al. (2015), Verbeke et al. (2017).

Taking the end product of this initial evolutionary stage, Mastropietro et al. (2021) study the evolution of these galaxies when injected on different orbits in the gaseous halo of a Fornax-like galaxy cluster. During this stage, filamentary structures form in

different orbital epochs. In our analysis we will consider a single temporal snapshot of a dwarf galaxy evolving on a generic orbit.

5.1.1. Extracted manifolds

We recover 15 streams of gas with varying lengths. In the following, for the sake of clarity, we discuss only the most elongated manifold recovered through the methodology. The manifolds are visualized via the two methodologies described in Sections 4.1 and 4.2, respectively. The values of the free parameters adopted for this analysis are shown in Table 9

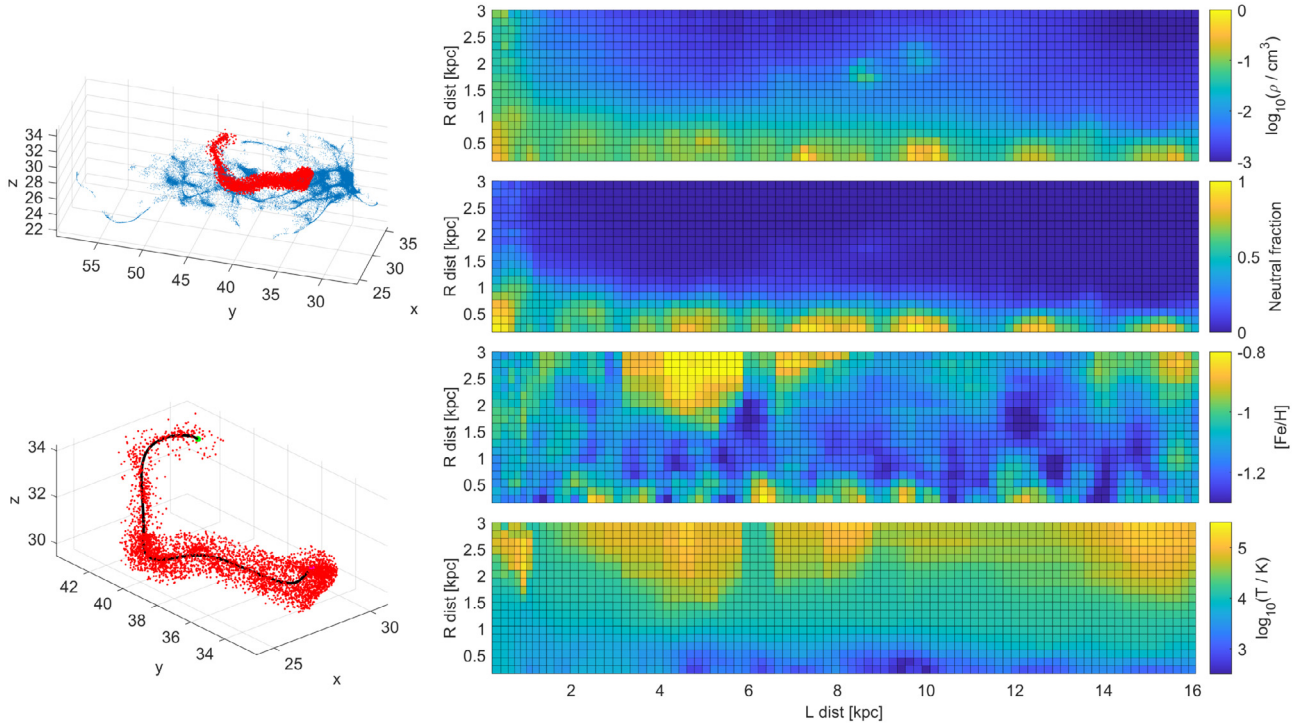


Fig. 16. Bi-dimensional profiles of one elongated manifold recovered via the methodology discussed in this work.

Table 9
Adopted values for experiments on Jellyfish galaxy.

LAAT	$r = 1$	$F_{th}^i = 5$
EM3A	$R_{min} = 0.5$	$R_{max} = 1.5$
Crawling	$r = 1$	
SGTM	$r = 1$	$S = L/2$
Bi-dim profile	$r_M = 3$	$c = 19$
Moving frames	$a = 2$	

Bi-dimensional profiles on jellyfish

Using the visualization tool described in Section 4.1 we present, in Fig. 16, one of the manifolds recovered via our methodology, departing from the head of the jellyfish and extending throughout its tail. Overall, while the elongation of individual manifolds varies, the inspected properties behave similarly for all detected structures. In particular, the neutral fraction (second bi-plot from the top), along with density (top bi-plot), is generally higher in the inner regions of the manifolds, across roughly their whole elongation. At the same time, temperature (bottom bi-plot, in logarithmic scale) is consistently lower in the same regions. A similar behavior cannot be found for the metallicity (third bi-plot from the top). It can be argued that this quantity is generally higher in the inner regions as well, however the non-uniformity of its distribution discourages an accurate inference on the causes of this anomaly. It is possible that, as the manifolds are from different regions of the galaxy, their interaction with the gas of the galaxy cluster's halo in previous epochs heavily influenced their evolution. Nonetheless, since the gas's metallicity does not directly affect the chances of its collapse, we can safely argue that the streams in jellyfish galaxies are effective loci of star formation. Furthermore, the core of these streams are more likely to contain newly born stars, and thus be observed via optical observations.

Co-moving orthonormal coordinate frames on jellyfish

The same quantities (ρ , neutral gas fraction, $[Fe/H]$, and T) have been studied via the Co-moving orthonormal coordinate frames technique and here presented for the same manifold

previously identified. The results are shown in Figs. 17 and 18 for four subsequent centers on the examined manifold. The three panels on top (left to right) and the one on the bottom left, show the distribution of these quantities on the perpendicular planes to the tangent bundle of the manifold. Over-plotted to the quantities distribution, in every panel we show the PDF of the probabilistic model obtained via SGTM, as (red) iso-curves. The red dot represents the center of the plane, corresponding to the embedded skeleton of the manifold. The bottom central and right panels show the location of the current plane with respect to the manifold and the whole data set respectively. For consistency, we will show in this work snapshots from the same manifold shown in Fig. 16.

Throughout the manifold's elongation we verify, via the snapshots presented in Figs. 17 and 18, the centering of the recovered skeleton with respect to the regions having the highest density. This region is also always associated to a higher neutral fraction and lower temperature. In the case of the studied manifold in agreement with the bi-dimensional profile (see Fig. 16), little can be said about the behavior of the metallicity, although there is a tendency of creating cores with local peaks surrounded by lower metallicity regions.

5.2. Kinematic study on cosmic web's filaments

On the many mega-parsec cosmological scales of the Universe, the spatial distribution of galaxies as well as clusters of galaxies is not uniform. In fact, looking at the results of the Sloan Digital Sky Survey (SDSS) (Fukugita, 1998; Gunn et al., 1998), one can see that there is an intricate, interconnected pattern that emerges at such a scale. This pattern forms a network now famously known as the cosmic web (Bond et al., 1996). As described in Peebles (1980), the cosmic web emerges as the outcome of the anisotropic nature of gravitational collapse. The latter is the driving force behind structure formation including the emergence of the cosmic web's different morphological components namely: clusters, filaments, and walls. The connection between these components

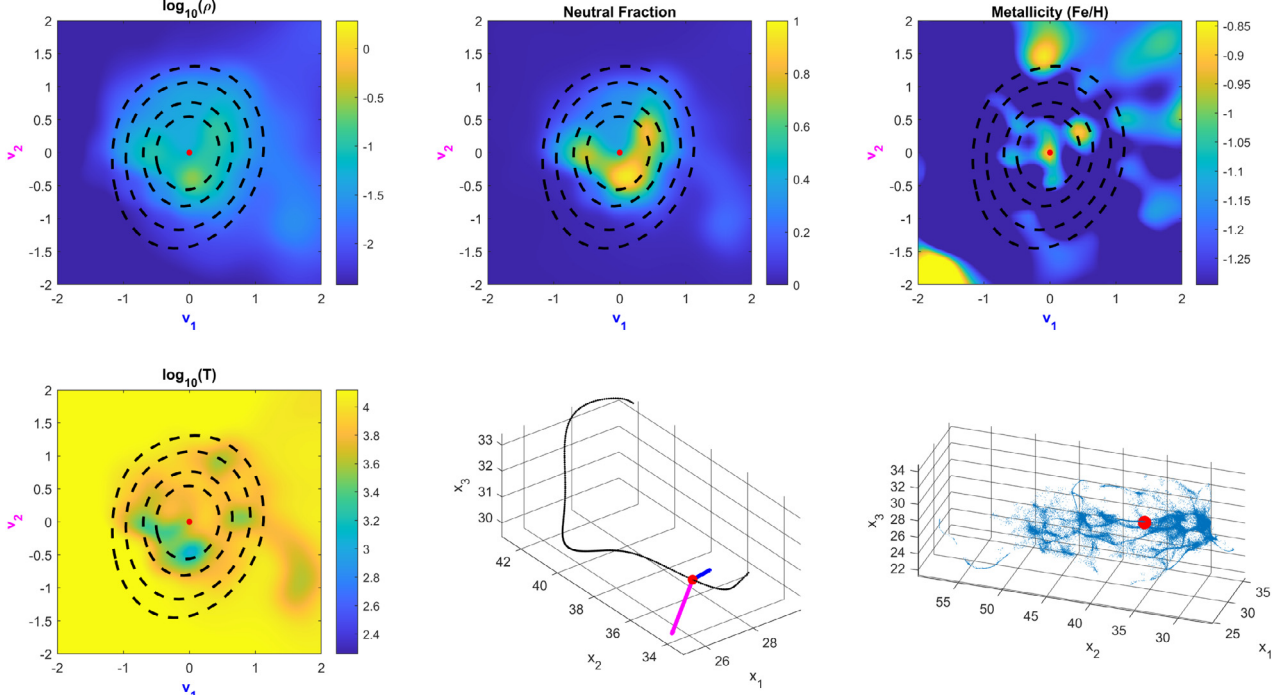
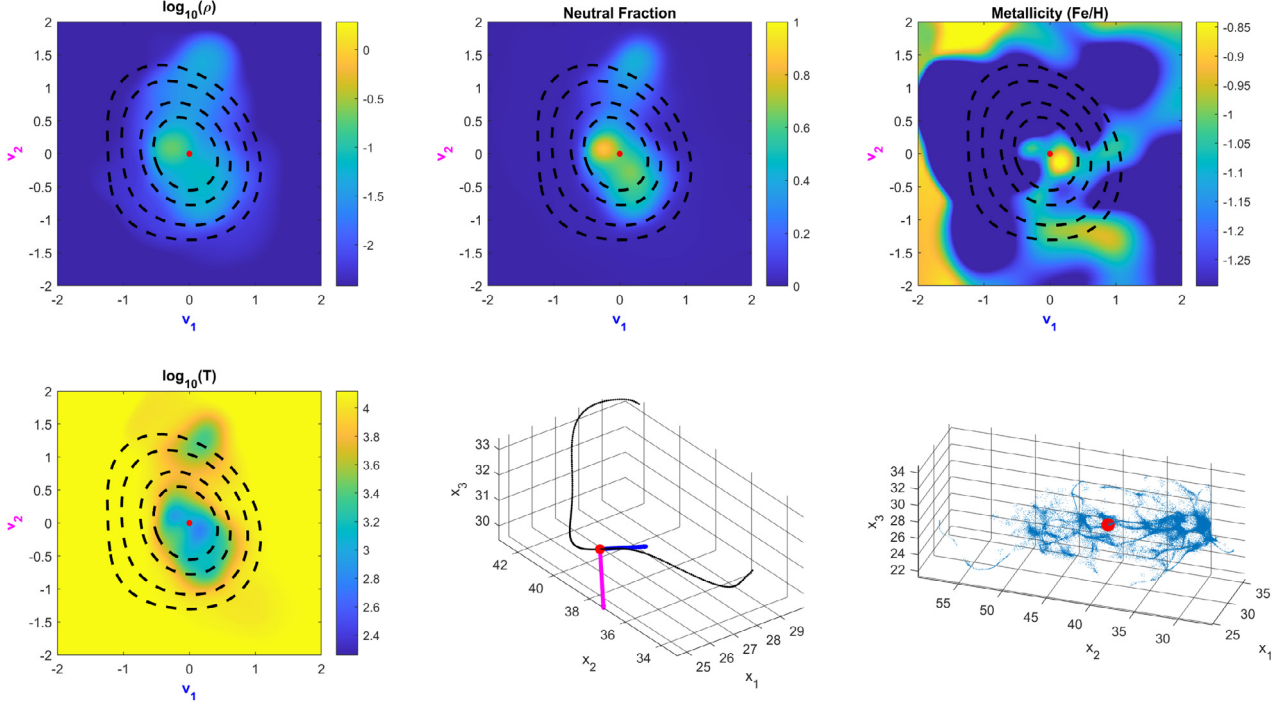
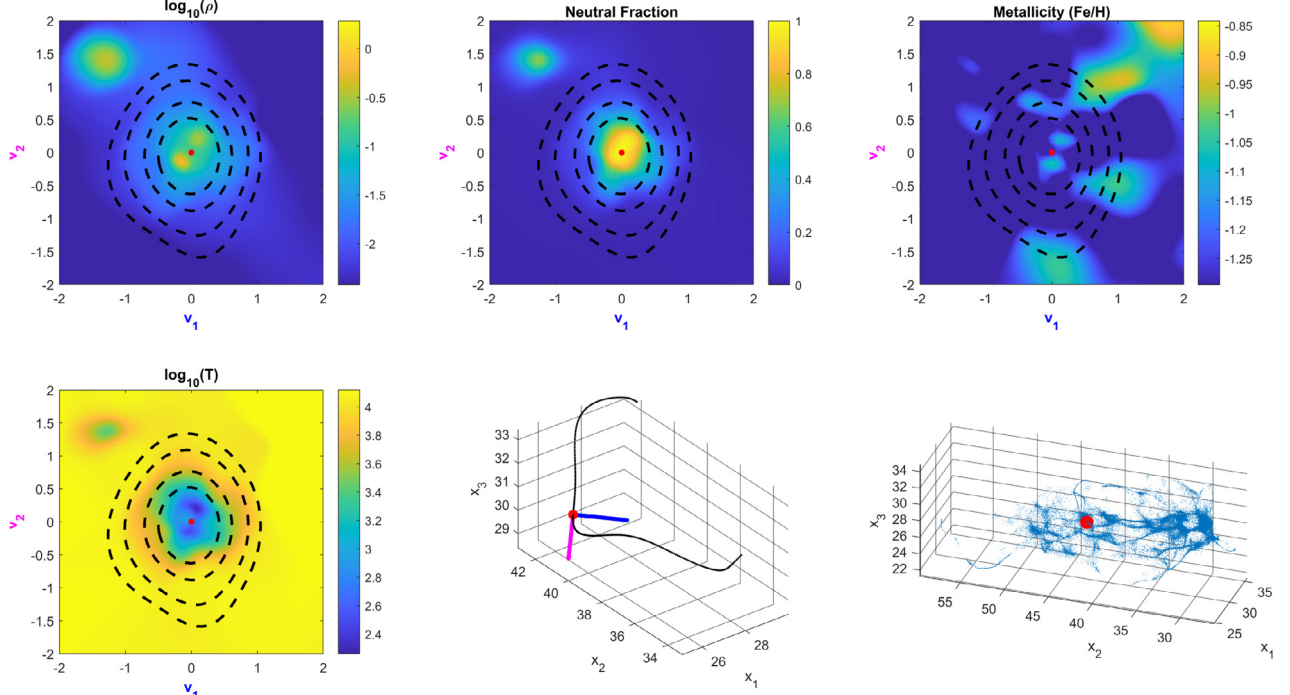
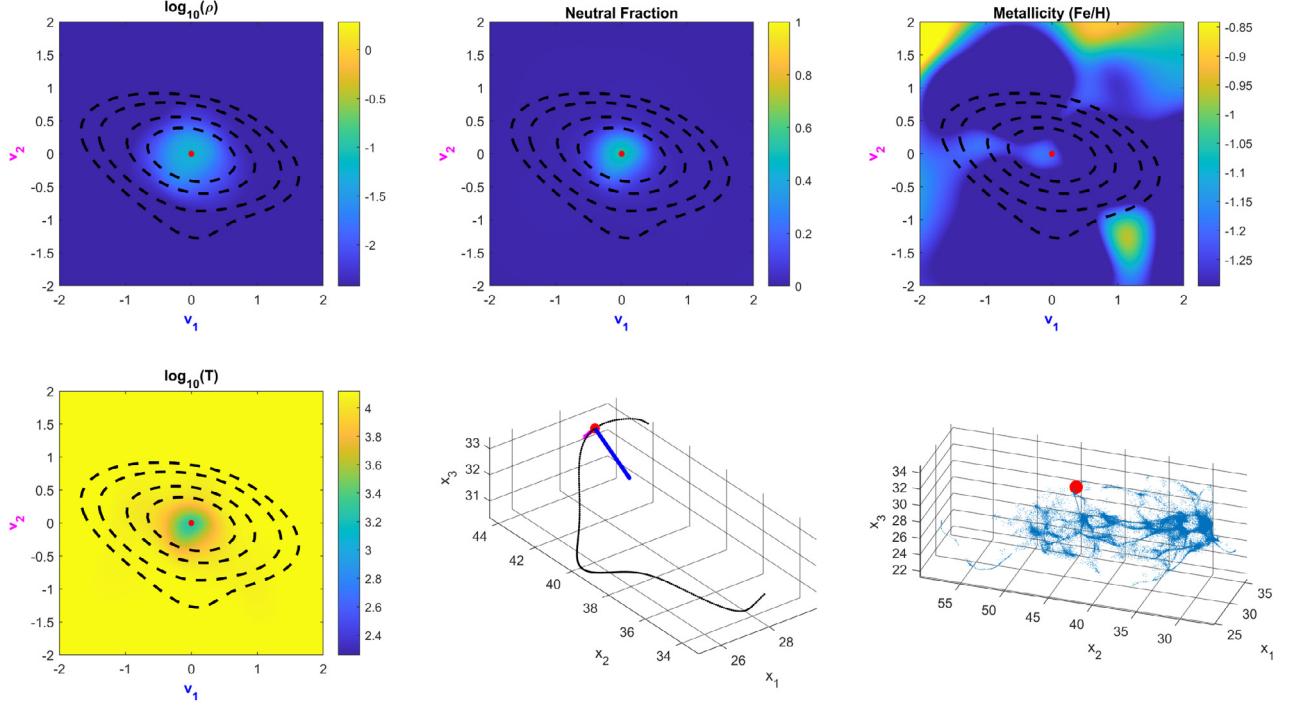
(a) Jellyfish manifold (Manifold \mathcal{M}_4 out of 15), snapshot n.34 of associated movie clip.(b) Jellyfish manifold (Manifold \mathcal{M}_4 out of 15), snapshot n.90 of associated movie clip.

Fig. 17. Equally spaced snapshots of the extracted manifold from jellyfish data set. Each panel in each sub-figure shows the distribution of a variable across the current orthonormal plane. The two bottom panels show the position of the current center (red sphere) on the detected manifold (black curve) and on the global diffused data set (blue dots, right panel).

can be summarized as follows: clusters are regions of intersection of filaments, and filaments are regions of intersection of walls (Doroshkevich, 1980; Shapiro et al., 1983; Pauls and Melott, 1995; Sathyaprakash et al., 1996; Cautun et al., 2013).

The growing interest in studying the cosmic web lies not only in its involvement in the cosmology domain, but also in its important influences on the evolution and properties of galaxies. For instance, in works such as Aragón-Calvo et al. (2007), Hahn et al. (2007a,b), and Paz et al. (2008) it has been shown that

the spin-orientation and shape of dark matter halos are distinctly influenced by the cosmic web environment they occupy (whether it is filamentary or sheetlike in nature). It has also been demonstrated that galaxies tend to have an alignment with the filaments that they inhabit (Jones et al., 2010; Tempel et al., 2013; Ganeshaiah Veena et al., 2018; Welker et al., 2019). Moreover, the influence of the cosmic web environment extends to other properties of galaxies such as their colors, gas content, and star formation rates (SFRs). Different studies have pointed out general

(a) Jellyfish manifold (Manifold \mathcal{M}_4 out of 15), snapshot n.136 of associated movie clip.**(b)** Jellyfish manifold (Manifold \mathcal{M}_4 out of 15), snapshot n.199 of associated movie clip.**Fig. 18.** Same as in Fig. 17 for additional centers on the manifold.

trends of these properties whereby galaxies closer to the cosmic web structures exhibit a lower specific SFR (are redder in color) and tend to be older, more metal rich and α -enhanced when compared to galaxies that have a larger distance to the structures (Rojas et al., 2004; Beygu et al., 2016; Chen et al., 2017; Kraljic et al., 2018; Winkel et al., 2021).

Given what has been presented, we illustrate the applicability of our methods by applying our pipeline on data sets pertaining to the cosmic web. The nature of the data is further described in the following section. We also reserve the detailing of the robustness

of the extracted and modeled morphologies to a second paper, where we narrow our focus to the cosmic web and the ability of our tool to trace out its different structures.

5.2.1. Generation of the cosmic web data set

We use a dark matter-only N -body cosmological simulation that was run using the GADGET-3 code. The initial conditions were generated at redshift $z = 200$ using the Multi Scale Initial Condition software (MUSIC; Hahn and Abel (2011)). The CAMB package (Code Anisotropies in the Microwave Background;

Table 10

Adopted values for experiments on Cosmic Web.

LAAT	$r = 1.5$	$F_{Th}^l = 5$
EM3A	$R_{min} = 1$	$R_{max} = 2$
Crawling	$r = 1.5$	
SGTM	$r = 1.5$	$S = L/2$
Moving frames	$a = 1$	

Lewis and Challinor (2011)) is used to calculate the linear power spectrum. We study a single cosmological volume with dimensions $120 \times 120 \times 120$ Mpc/h. The dark matter (DM) particles have a fixed mass of $1.072 \times 10^9 M_\odot/h$, and a cosmology of $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_b = 0.047$ and $h_0 = 0.684$ was assumed for the initial conditions and for the simulation itself. In this project, we consider the redshift zero output file, which consists of the masses, velocities and positions of all the dark matter particles in the present.

5.2.2. Extracted manifolds

Methodologies: LAAT \rightarrow EM3 A \rightarrow Dimensionality Index \rightarrow Crawling \rightarrow SGTM \rightarrow Co-moving Orthonormal coordinate frames.

Adopting the methodology presented previously, we isolate two manifolds in the simulated volume, connected by a node of the Cosmic Web. By building orthonormal co-moving frames over the skeletons of these manifolds, we study their DM distribution and kinematic properties. The values of the free parameters adopted for this analysis are shown in Table 10. For each particle in the simulation, at each location over the manifold's skeleton, we compute its tangential and orthogonal velocity with respect to the local reference frame. We look at the DM particles distribution as a discrete sample of the actual dark matter, where each particle is representative of the kinematic state of a fixed size neighborhood.

Given a specific point $\tilde{\mathbf{t}}_\ell \in \overline{\mathcal{P}}_t^k$ belonging to manifold \mathcal{M}_k 's up-sampled skeleton, as a product of our methodology we recover the corresponding tangent vector $\hat{\xi}_\ell$ and perpendicular plane $\mathcal{T}_\ell^\perp = \text{span}\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2\}$. As discussed in Section 4.2, we obtain the index set \mathcal{J} of all particles falling within the box of size $\bar{d}_\ell \times 2b$ centered on $\tilde{\mathbf{t}}_\ell$. Each \mathbf{t}_m such that $m \in \mathcal{J}$, has a velocity $\mathbf{v}_m \in \mathbb{R}^3$ describing its motion. The velocity vector is projected onto \mathcal{T}_ℓ^\perp and along the tangent vector $\hat{\xi}_\ell$ to manifold \mathcal{M}_k on $\tilde{\mathbf{t}}_\ell$:

$$\mathbf{v}_m^\perp = \mathbf{P} \mathbf{v}_m; \quad \mathbf{v}_m^\parallel = (\mathbf{v}_m \cdot \hat{\xi}_\ell) \hat{\xi}_\ell, \quad (63)$$

for every $m \in \mathcal{J}$. We can now compute the weighted mean of the two projected velocities over points on \mathcal{T}_ℓ^\perp :

$$\bar{\mathbf{v}}(\mathbf{y}_{ij})^\perp = \frac{\sum_{m \in \mathcal{J}} p(\mathbf{t}_m^\perp | \mathbf{y}_{ij}, \delta) \mathbf{v}_m^\perp}{\sum_{q \in \mathcal{J}} p(\mathbf{t}_q^\perp | \mathbf{y}_{ij}, \delta)} \quad (64)$$

$$\bar{\mathbf{v}}(\mathbf{y}_{ij})^\parallel = \frac{\sum_{m \in \mathcal{J}} p(\mathbf{t}_m^\perp | \mathbf{y}_{ij}, \delta) \mathbf{v}_m^\parallel}{\sum_{q \in \mathcal{J}} p(\mathbf{t}_q^\perp | \mathbf{y}_{ij}, \delta)}, \quad (65)$$

where $p(\mathbf{t}_m^\perp | \mathbf{y}_{ij}, \delta)$ is the Gaussian kernel defined in Eq. (61). Note that parameter δ can be fixed by the user to match a desired smoothing, however in our experiments we fix it to $\delta = \frac{\sqrt{a/M}}{4}$: the half-diagonal of the square formed by four adjacent points sampling \mathcal{T}_ℓ^\perp in \mathcal{V}_ℓ . Having obtained a velocity field on \mathcal{T}_ℓ^\perp , we are now able to compute its rotor and divergence with respect to the local coordinate frame. We define the ∇ operator on the local reference frame given by $\text{span}\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\xi}_\ell\}$ as $\nabla = (\partial/\partial u_1, \partial/\partial u_2, \partial/\partial \xi_\ell)$, so that the rotor and divergence of vector field $\bar{\mathbf{v}}$ can be written as

$$\nabla \times \bar{\mathbf{v}}^\perp = \left(\frac{\partial \bar{\mathbf{v}}^{\perp,3}}{\partial u_1} - \frac{\partial \bar{\mathbf{v}}^{\perp,1}}{\partial \xi_\ell} \right) \hat{\mathbf{u}}_1$$

$$+ \left(\frac{\partial \bar{\mathbf{v}}^{\perp,1}}{\partial \xi_\ell} - \frac{\partial \bar{\mathbf{v}}^{\perp,3}}{\partial u_1} \right) \hat{\mathbf{u}}_2$$

$$+ \left(\frac{\partial \bar{\mathbf{v}}^{\perp,2}}{\partial u_1} - \frac{\partial \bar{\mathbf{v}}^{\perp,1}}{\partial u_2} \right) \hat{\xi}_\ell$$

$$= \left(\frac{\partial \bar{\mathbf{v}}^{\perp,2}}{\partial u_1} - \frac{\partial \bar{\mathbf{v}}^{\perp,1}}{\partial u_2} \right) \hat{\xi}_\ell; \quad (66)$$

$$\nabla \cdot \bar{\mathbf{v}}^\perp = \frac{\partial \bar{\mathbf{v}}^{\perp,1}}{\partial u_1} + \frac{\partial \bar{\mathbf{v}}^{\perp,2}}{\partial u_2} + \frac{\partial \bar{\mathbf{v}}^{\perp,3}}{\partial \xi_\ell} = \frac{\partial \bar{\mathbf{v}}^{\perp,1}}{\partial u_1} + \frac{\partial \bar{\mathbf{v}}^{\perp,2}}{\partial u_2}. \quad (67)$$

The final forms of the rotor and divergence are obtained by noting that only the third term in the expansion is non-null in the first case, while it is the only null element in the second (being the vector field over the plane).

The results for one manifold selected from the simulated volume are shown in Figs. 19–20.

In each sub-figure, information is presented by the following scheme:

Skeleton (solid, black line) recovered by up-sampling the embedded graph obtained via SGTM and isosurface of the corresponding model's PDF (red, opaque surface). Extracted simulated particles obtained via soft assignment on the whole data set, given the model (gray dots). Current center $\tilde{\mathbf{t}}_\ell$ (red sphere) of the perpendicular plane \mathcal{T}_ℓ^\perp and vectors $\hat{\mathbf{u}}_1$ (blue arrow) $\hat{\mathbf{u}}_2$ (magenta arrow) spanning it (top left panel). Heatmap of the weighted mean projected tangential velocity $\|\bar{\mathbf{v}}^\parallel\|$, at rest frame: the velocity of the central point $\bar{\mathbf{v}}^\parallel(\tilde{\mathbf{t}}_\ell)$ has been removed in order to represent local velocities with respect to the local frame. Over-plotted are the iso-contours of the model's PDF (dashed black lines) and a mask hiding sparsely populated regions of plane \mathcal{T}_ℓ^\perp (top middle panel).

Quiver plot (vector plot) of the weighted mean velocity field (black arrows) over the plane with over-plotted iso-contours of the model's PDF. Again the central velocity $\bar{\mathbf{v}}^\perp(\tilde{\mathbf{t}}_\ell)$ has been subtracted to the velocity field (top right panel).

Number of particles within the sphere of radius κ for each point $\mathbf{y}_{ij} \in \mathcal{V}_\ell$, sampling regularly plane \mathcal{T}_ℓ^\perp , with over-plotted mask (bottom left panel).

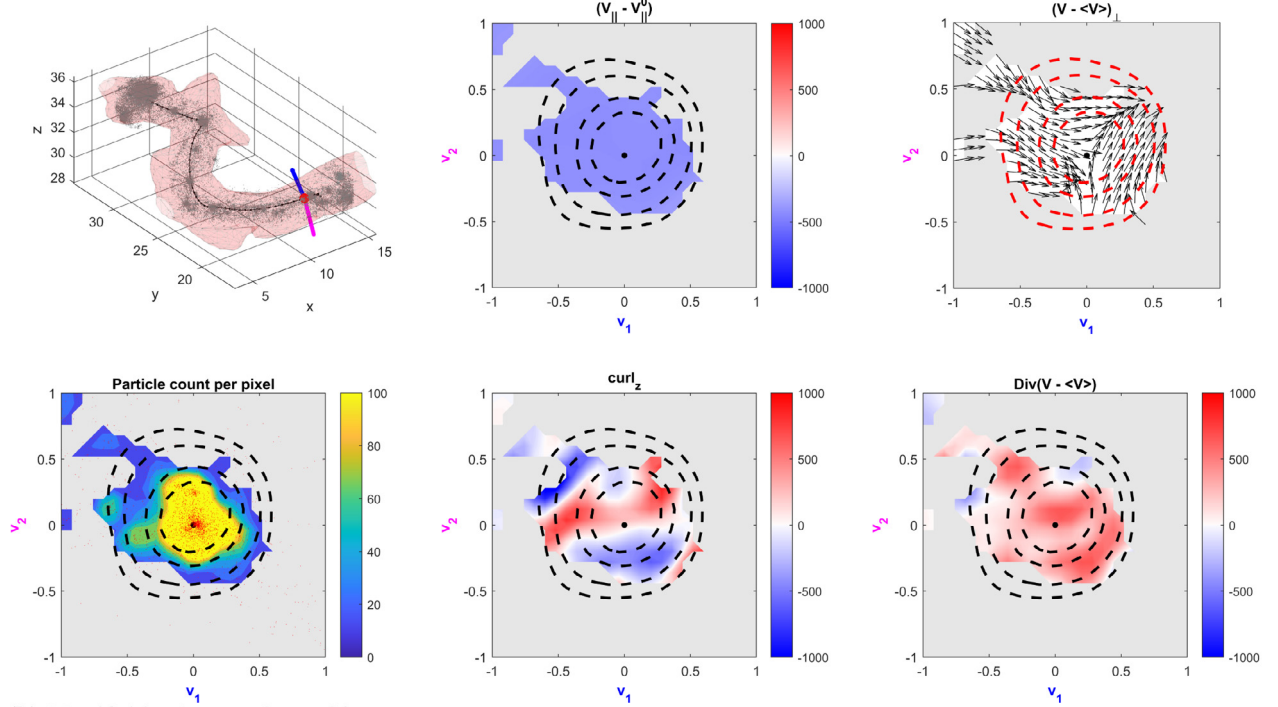
Tangential component of the rotor (also called “curl”) of the weighted mean velocity field computed over the perpendicular plane (bottom middle panel).

Divergence of the weighted mean velocity field computed over the perpendicular plane (bottom right panel).

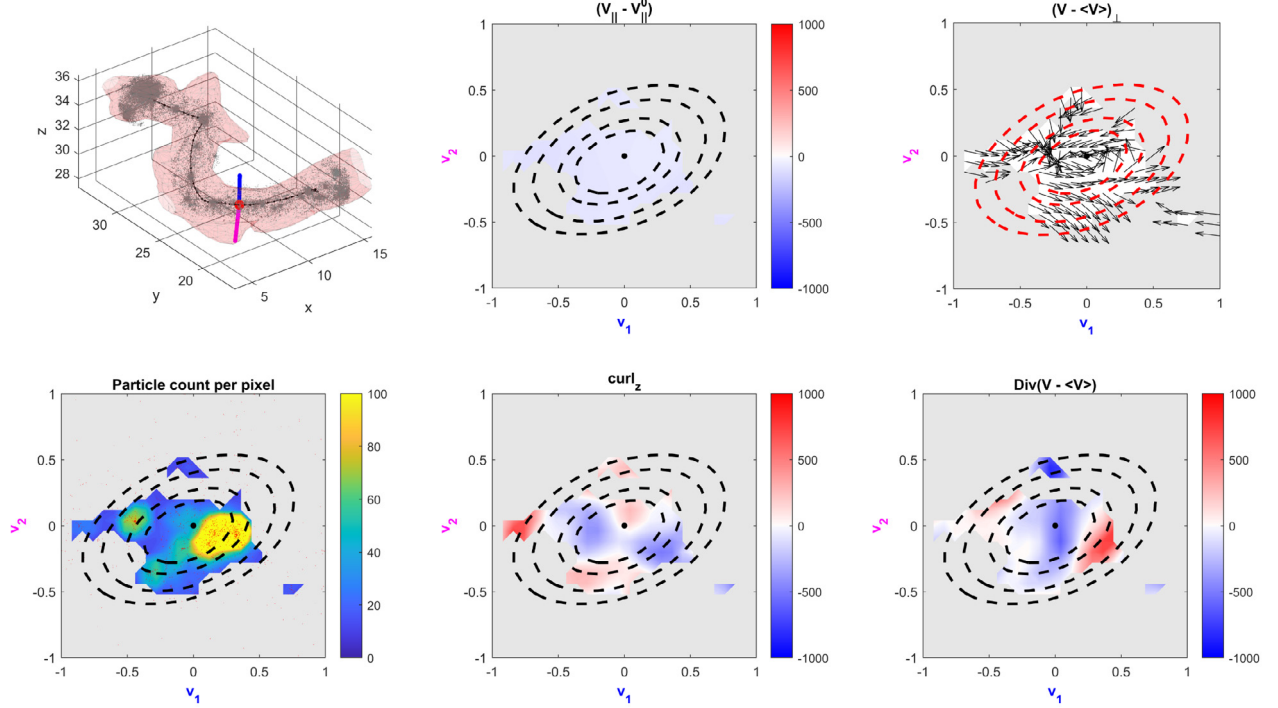
A few main results can be drawn from the visualization of the kinematic properties of the two manifolds using our technique:

- The skeletons of the manifolds are generally aligned with their corresponding densest regions, when these are unique. In cases where mass has a multi-modal distribution, the center of the plane is usually placed in order to include all modes. This is visible in the bottom left panel of each figure. The regions containing the largest amount of particles (and thus the highest density) are usually centered on the plane when presenting one peak, or slightly shifted when more than one peak can be found.
- The tangential velocity of particles on the manifold (top central panel) has the tendency to change sign at a certain distance from the two extremities. The extremities of each manifold are the nodes of the cosmic web (i.e. the clusters). In particular, starting from one node and moving towards the other, the tangential velocity has a negative sign at first (overall blue color), meaning that particles are pulled towards the starting node. As we move towards the second node, the tangential velocity's module tends to decrease until it reaches zero. After this “saddle” point, the tangential velocity is aligned with the crawling direction, meaning that

(a) Manifold n.1, snapshot n.9



(b) Manifold n.1, snapshot n.20

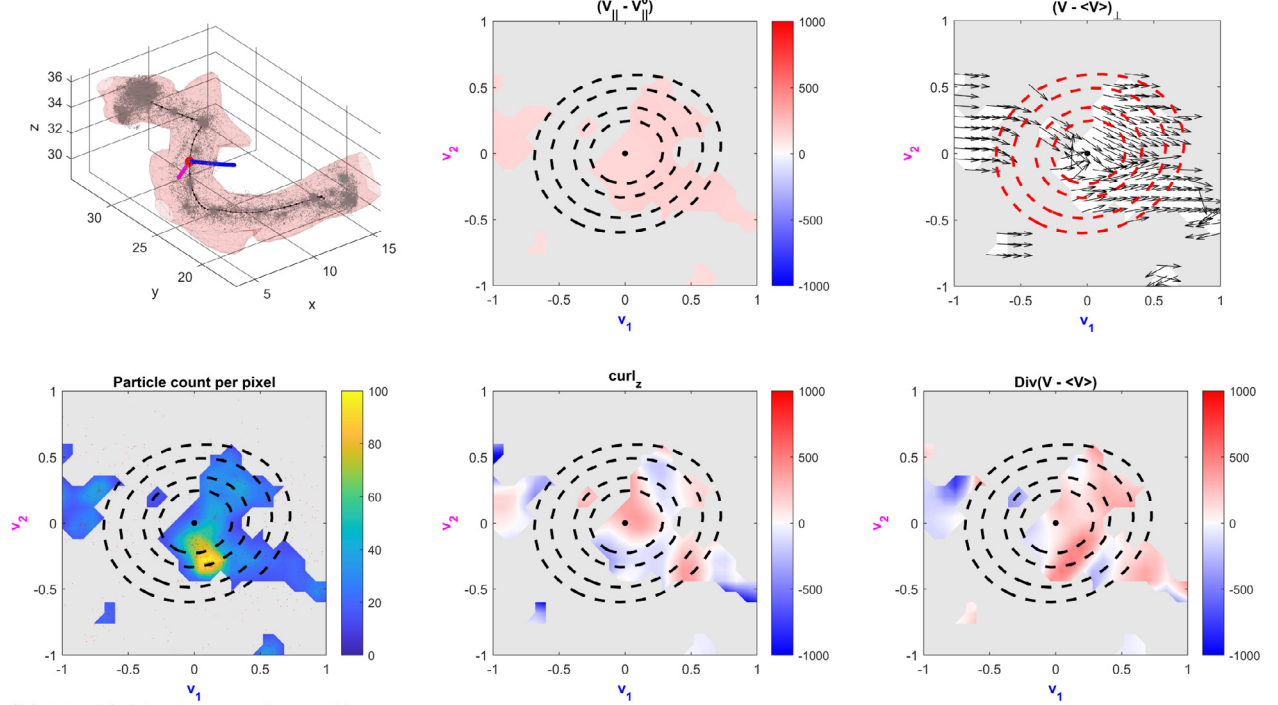
**Fig. 19.** Different snapshots of orthonormal coordinate frame movie for manifold n.1 of the Cosmic Web.

the particles tend to be pushed towards this second node (the parallel velocity map turns red). Fig. 20a clearly shows the flip in sign of the tangential velocity. Fig. 20b, shows a slight deviation from this behavior. This is due to the presence of a mass concentration within the filament (also identifiable in the bottom left panel of the figure) locally pulling particles towards itself.

- The parallel curl (perpendicular to the plane) of the velocity field shows clear islands of opposite signs. This result is in agreement with what found in Laigle et al. (2014), although

the analysis is only performed on one filament here. In their work, the number of regions with opposite curl shows a large variability throughout the elongation of the stream. It is however also pointed out that four regions are most commonly found. In our analysis we generally agree with the findings of Laigle et al. (2014) in terms of the variability in the number of these regions, but further analysis is needed to confirm the most common occurrence. However, in some iterations, we are in agreement with their prediction (e.g. Fig. 19b). Future studies will focus on applying

(a) Manifold n.1, snapshot n.39



(b) Manifold n.1, snapshot n.50

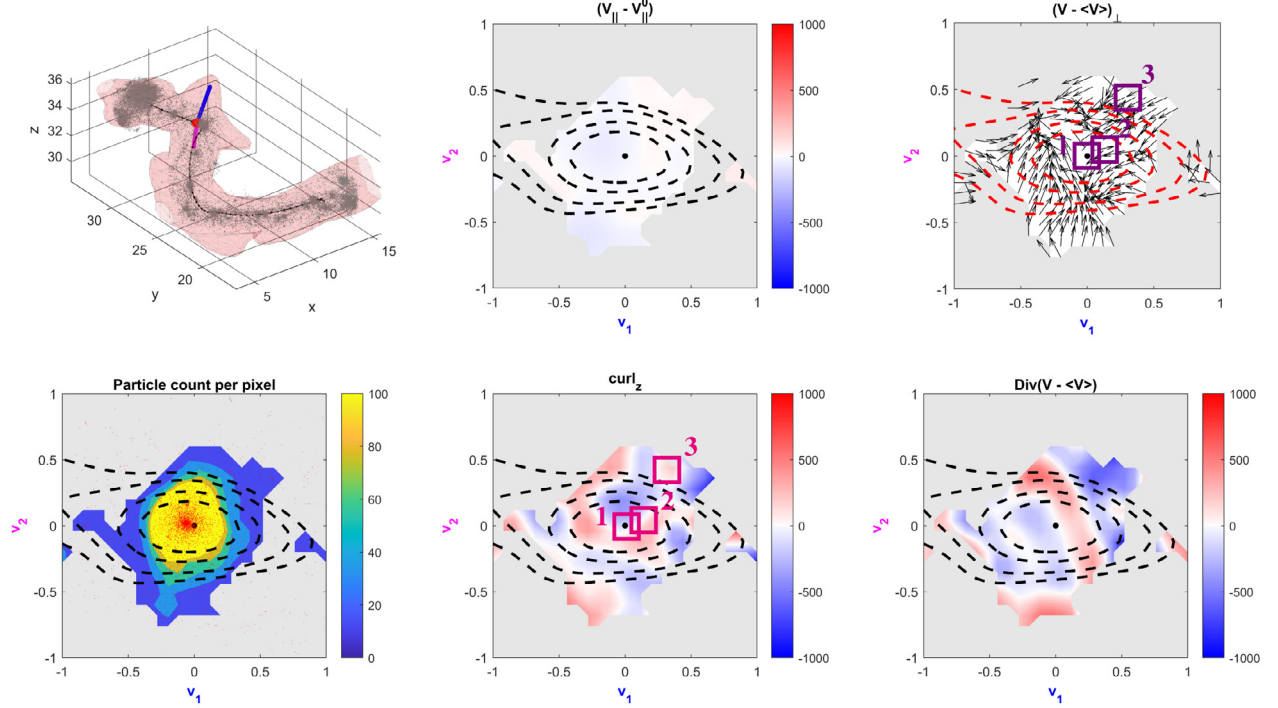


Fig. 20. Same as in Fig. 19 for further centers on the manifold. The bottom central panel (curl) and top right panel (velocity field) of Fig. 20 also present numbered square regions in shades of purple. The three numbered squares identify regions presenting opposite sign in curl and their respective velocity field. Zoomed in pictures of the velocity field in the three regions are presented in Fig. 21.

the proposed methodology to a larger set of filaments in order to obtain a robust estimate about the variability in the number of these regions across a larger population.

- Particles in the outskirts of the manifolds are attracted towards their cores. This result transpires from a visual inspection of the top right panels of all figures. The weighted mean perpendicular velocity field is generally oriented towards the center of the plane (the manifold's core).

- In Fig. 20b we identify three regions of high vorticity from the bottom central panel. The three square regions are numbered and shown correspondingly on the top right panel. The same regions are presented in Fig. 21a–c respectively. The velocity map in the three regions confirms the behavior presented in the colored regions in bottom central panel of Fig. 20b. In particular, the “blue-curl” region 1 presents a

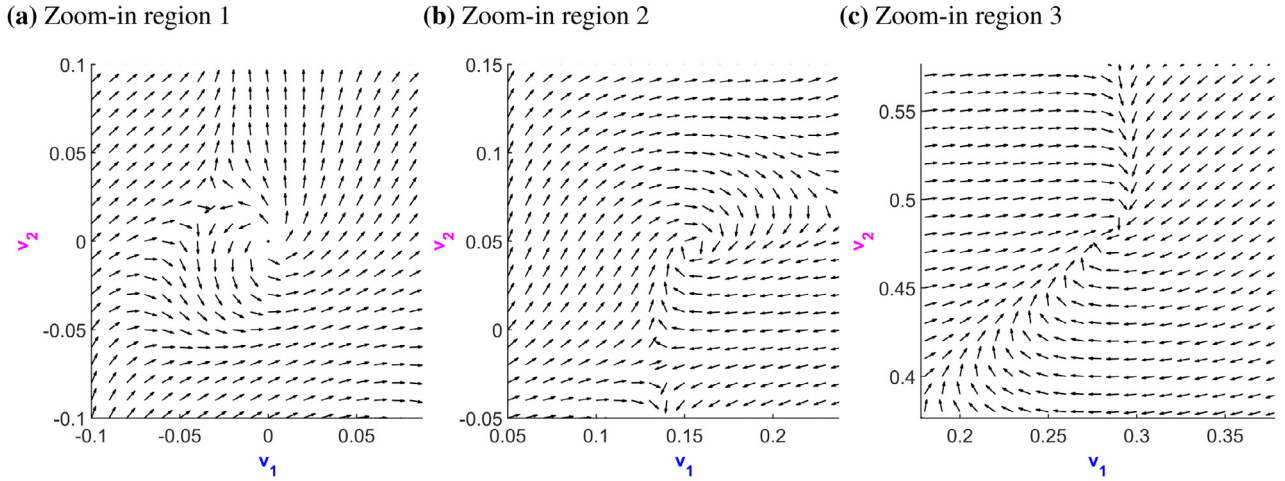


Fig. 21. Zoomed-in regions of high vorticity identified in Fig. 20b. Note how local vortices can be identified within each panel. The corresponding values for the curl can be identified in the bottom central panel in Fig. 20b. The zoom-in region 1 curl (blue) is opposite to zoom-in 2 and 3 (red) and this is reflected in the orientation of the vortices: counter clockwise for 1 and clockwise for 2 and 3.

vortex oriented counter clockwise, while the vortices in the “red-curl” regions 2 and 3 are oriented clockwise.

To summarize, our tool was applied to a simulated cosmic web volume, and it was able to recover from the complex point clouds streams of complicated morphology. The proposed analysis of the structures is in agreement with previous findings regarding the dynamics of the filaments. This example of application of our toolbox is further proof of the validity of the proposed methodologies. Furthermore, it is still possible to extend the analysis to other non-dynamical properties of the cosmic filaments (e.g. as shown in Section 5.1 for the jellyfish galaxy), as well as their morphology, with no further effort in the development of new tools. As this will be subject of future studies, we believe that the toolbox here presented and demonstrated may prove extremely useful in the understanding of these structures.

5.3. ω -Centauri's stellar stream from GAIA-DR2

The study of the stellar galactic halo of the Milky Way is of great importance to astronomers interested in the archaeological aspects of the Galaxy, particularly because of the halo's key role in characterizing the galaxy's formation history. A crucial component to the formation of galaxies in a hierarchical formation scenario is the growth by tidal disruption or mergers with external astronomical objects. This led to the deposit of merger debris in the Milky Way's halo in the form of stellar streams or stellar overdensities (Helmi, 2020). In order to characterize the interaction history of the Milky Way, astronomers have tracked the stars found within stellar streams in the halo allowing them to trace the stars' origins back to the early phases of the Galaxy's formation (Helmi, 2020).

Moreover, dynamically cold stellar streams provide an opportunity to probe the acceleration field of the Galaxy both locally and globally (Johnston, 1999; Ibata et al., 2002; Johnston et al., 2002; Carlberg, 2012). This gives great insight onto the nature of the gravitational force and the distribution of dark matter both of which are encoded in the Milky Way's acceleration field (Ibata et al., 2021).

Deep wide-field photometric surveys including the SDSS (York et al., 2000), PanSTARRS (Chambers and Pan-STARRS Team, 2016), and DES (Abbott et al., 2018) have increased our knowledge of the Galaxy's stellar halo by revealing many of the narrow streams and overdensities belonging to it (Helmi, 2020). However, much greater clarity was obtained with the Gaia mission following the

second Gaia data release (DR2; Gaia Collaboration, 2018) and the third early data release (EDR3; Gaia Collaboration, 2020).

Given the multidimensional data provided in Gaia EDR3, the Milky Way's stellar streams constitute another natural application of our pipeline. For a test subject, we have chosen ω -Centauri as it is the largest cluster known and has been extensively proven to be tidally disrupted. In particular we base our work on the information provided in Ibata et al. (2019a) for spotting the tidal arms of ω -Centauri. Further detailing of this procedure will be provided in Section 5.3.1. Through this third demonstration, we show that our modeling is applicable not only to simulation outputs, but also to the ever-increasing amounts of observational data.

5.3.1. Isolation of ω -Centauri's stream

Methodologies: EM3 A \rightarrow Dimensionality Index \rightarrow Crawling \rightarrow SGTm \rightarrow Co-moving Orthonormal coordinate frames (mod.).

In this section, we outline the different steps followed to spot the tidal-arms of Omega-Centauri (ω -Cen). This will act as the preprocessing stage before applying our algorithms to extract and model the targeted streams. Through the N -body simulations conducted in Ibata et al. (2019a), the “Fimbulthul structure” (Ibata et al., 2019b) in Gaia DR2 was identified as part of the tidal arm of the cluster. The properties of the system that were guided by the results of their N -body simulations then served as a selection filter applied to the stars in an area around the cluster. In order to obtain a distribution of the stars that show the two streams, we follow their selection criteria and refer the reader to Ibata et al. (2019a) for a detailed motivation of the selection basis.

From the Gaia archive, we choose a rectangular region spanning $l = [-70^\circ, -30^\circ]$ and $b = [5^\circ, 50^\circ]$ where l and b are the galactic longitude and latitude respectively. In this region, we select the stars that have a parallax uncertainty less than 1 mas and those with parallax measurements consistent within 1σ with distances between 4 and 6 kpc.

We then apply a filter on the kinematic behavior where we select the stars that have proper motions along the declination direction μ_δ similar to that of the cluster, and proper motions along the right ascension direction μ_α that show a decreasing linear gradient as a function of b . The rate of decrease is taken as 0.125 mas/yr for every degree in b . For these requirements, $\mu_\alpha = -3.1925 \pm 0.0022$ mas/yr and $\mu_\delta = -6.7445 \pm 0.0019$ mas/yr are chosen as the reference values for the cluster, and the stars within 1 mas/yr of the two kinematic criteria are selected.

Table 11Adopted values for experiments on ω -Centauri.

EM3A	$R_{min} = 1$	$R_{max} = 2$
Crawling	$r = 1.5$	
SGTM	$r = 1.5$	$S = L/2$
Moving frames	$a = 2.5$	

Furthermore, to correct for interstellar extinction, we use the dust maps provided in Schlegel et al. (1998) and recalibrated by Schlafly and Finkbeiner (2011) to modify the brightness and colors of the remaining stars. The extinction-corrected magnitudes are obtained assuming foreground-only interstellar extinction with $R_V = 3.1$. For choosing the stars belonging to the Color-Magnitude Diagram (CMD) of the cluster, we draw the polygons shown in Figure 4a in Ibata et al. (2019a) and select the stars belonging to the regions within those polygons. This selection rule allows for the filtering out of a large number of background stars while minimizing the bias against selecting stars belonging to ω -Centauri (Ibata et al., 2019a).

5.3.2. Modeling of stream via SGTM

All the selected particles are collected in data set \mathcal{Q} . We first apply EM3A 3.1.2. In fact, when removing particles within a circular region centered on the core of ω_{Cen} , the point distribution is irregular in its vicinity, causing a sharp discontinuity between the dense (Fig. 22a, lower left) and sparse (Fig. 22b, upper right) filaments. Furthermore, the over-density of the galaxy on the bottom left region of the panel, has a higher chance of pheromone being deposited there than on the filaments. This results in LAAT not being able to equally distribute pheromone along the regions of interest. However, the application of EM3 A to the data set obtained by kinematical and color-magnitude filtering proved successful, being EM3 A less influenced by global densities and more focused on local anisotropies. Since the data set has large variations of the local densities, particular care has to be taken when applying the methodology. In particular, adopting a large number of iterations with a large radius for neighborhood search may result in the production of spurious structures, unrelated to the filament we require to extract. On the other hand, a small radius may disrupt the main structure of interest prematurely, fragmenting it in multiple clumpy regions. It is thus advisable to monitor the advancement of the methodology at each iteration and to test different values for its hyper-parameters. Furthermore, since ω_{Cen} 's main body has the highest star's number density, we remove from data set \mathcal{Q} all the stars lying within the sphere of radius $r = 0.8$ deg, centered at [309.10202, +14.96833] in the galactic coordinate reference system. Following the application of the proposed methodologies, we recover the skeleton and SGTM model of ω_{Cen} 's stream. The resulting SGTM is shown in Fig. 22b, together with the stars selected by the criteria in Section 5.3.1 (Fig. 22a), as obtained by Ibata et al. (2001). The parameters used for the methodologies are shown in Table 11. The red line is the skeleton after training SGTM, the noise model is represented by the isocurves at different isovalues of the model's likelihood over the data space. The stream's model resembles very closely the one highlighted in Fig. 22a. Gray dots in Fig. 22b are the same points presented in Fig. 22a, however we omit the coloring in order to enhance visibility of the noise model's iso-contours. The stream is analyzed via the methodology presented in Section 4.2, but adapted for the 1D case.

Co-moving orthonormal reference frame

We now consider the co-moving orthonormal reference frame technique, for this one-dimensional case. Figs. 23a–c show the results on ω_{Cen} 's streams. The yellow strip on the top-right panel

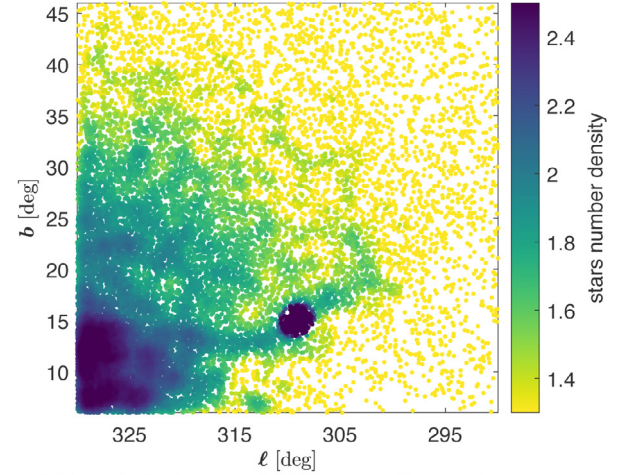
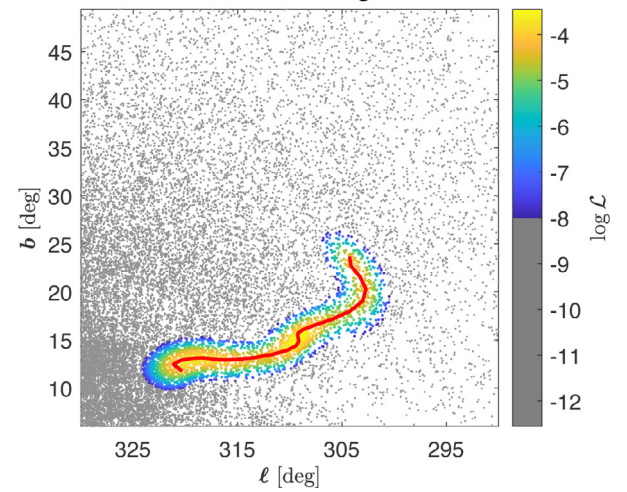
(a) GAIA DR2 stars (adapted from Ibata et al. (2001))**(b) Probabilistic model of Omega-Centauri stream**

Fig. 22. Panel (a) depicts the GAIA DR2 stars after selection by criteria explained in Section 5.3.1, colored by local stellar log-density (adapted from Ibata et al., 2001). Panel (b) shows the probabilistic model of Omega-Centauri's stream as obtained from crawling and SGTM. The red line marks the trained skeleton and the stars are colored by the log likelihood with a threshold to gray to visualize isocurves.

of each figure, shows the selection of stars used for computing local density of the stream. The selection is made on a geometrical basis, by projecting all stars onto the tangential and perpendicular spaces of the manifold at each node in the trained SGTM. For each point in $\tilde{\mathcal{P}} \in \overline{\mathcal{P}}$ (we drop index k , since we only consider one manifold), we know the tangent vector to the manifold at $\tilde{\mathbf{t}}_\ell$: $\tilde{\xi}_\ell = [\xi_\ell^1, \xi_\ell^2]$. Its orthogonal complement is given by either $\tilde{\xi}_\ell^{\perp+} = [-\xi_\ell^2, \xi_\ell^1]$ or $\tilde{\xi}_\ell^{\perp-} = [\xi_\ell^2, -\xi_\ell^1]$, giving the vectors pointing towards increasing or decreasing b respectively. We chose here to adopt $\tilde{\xi}_\ell^{\perp+} = \xi_\ell^{\perp+}$ as the orthogonal complement to tangent vector $\tilde{\xi}_\ell$.

At each $\tilde{\mathbf{t}}_\ell$, we project all stars in \mathcal{Q} onto the tangent vector $\tilde{\xi}_\ell$. To define the width of the selection strip, we impose a maximum module for the orthogonal projection onto $\tilde{\xi}_\ell$. In this case the maximum tangential distance is fixed at the distance between two adjacent centers in the SGTM. We then project all stars onto $\tilde{\xi}_\ell$'s orthogonal complement $\tilde{\xi}_\ell^{\perp+}$ and impose a maximum perpendicular distance of 8° . The stars satisfying these two restrictions are plotted in the central panel of each plot (yellow points) in Fig. 23 and it can be verified that they form indeed a rectangular shape with shorter edge aligned with the tangential space to

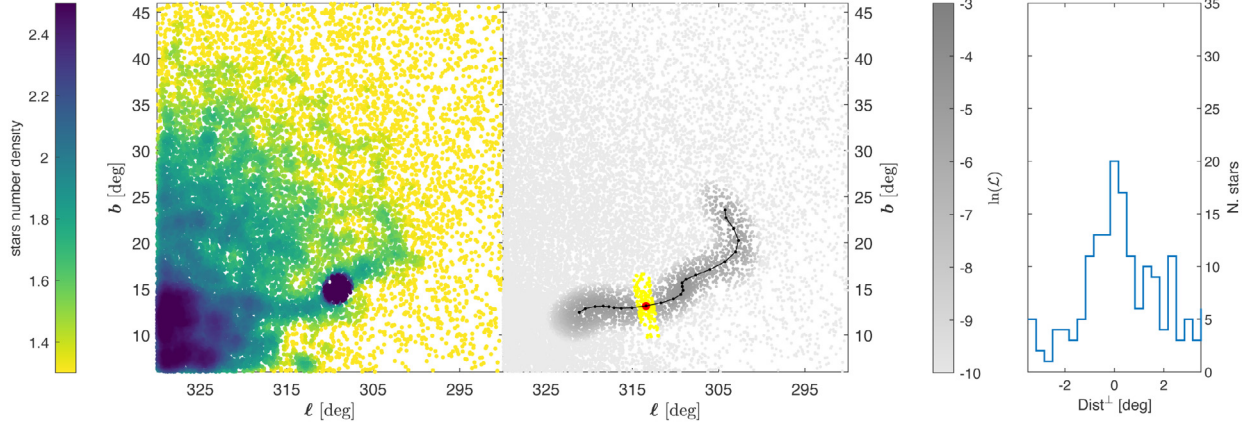
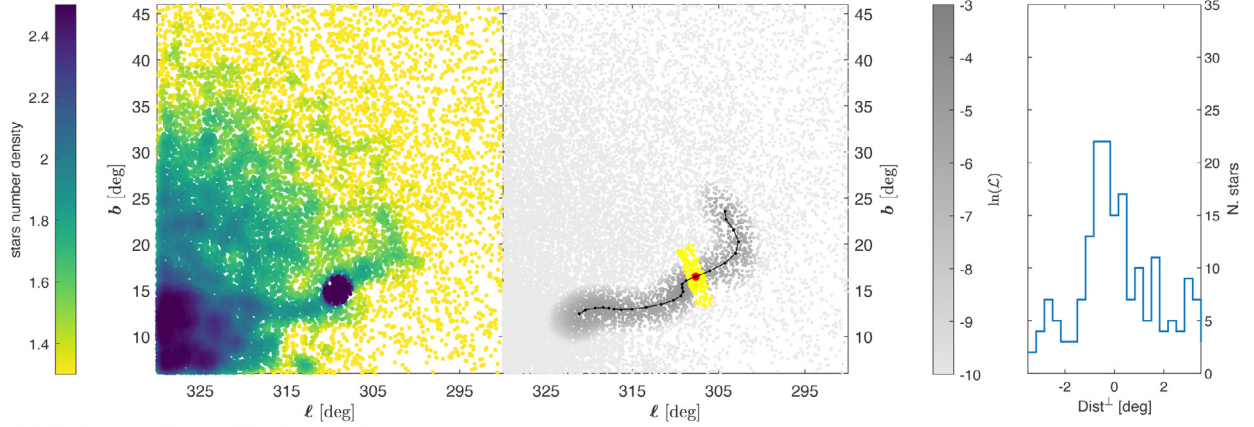
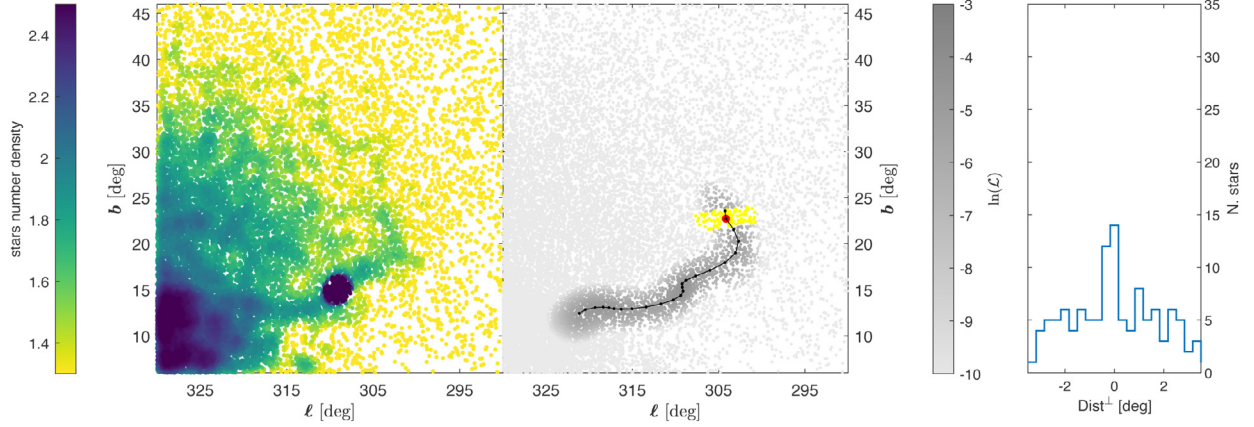
(a) Reference frame 9 of ω_{Cen} stream(b) Reference frame 17 of ω_{Cen} stream(c) Reference frame 23 of ω_{Cen} stream

Fig. 23. Visualizations of ω_{Cen} 's stream recovered from GAIA DR2. Each row contains the original density map (left), the model likelihood with the visualization frame position marked by a red dot (middle). The yellow window expands perpendicular to the stream and can be used to inspect properties at different positions by analyzing sources within, such as distribution perpendicular to the stream (right histogram). (a) and (b) show regions where the stream peaks clearly, while the reference frame in panel (c) exemplifies a fainter stream part exhibiting a flatter histogram.

the manifold and the longer edge aligned with its orthogonal complement. We then partition the orthogonal axis in bins and form a histogram of the number of particles per bin, as shown in the right panels of Fig. 23, at varying $\tilde{t}_\ell \in \overline{\mathcal{P}}$. The histogram shows how the stellar number density is generally higher nearby the middle of the perpendicular axis, where the current center of SGTm is located.

In order to estimate the detection quality of the stream we evaluate the Signal-to-noise ratio (SNR) of the central counts with respect to a co-moving background. For each point on the stream, we define three apertures of fixed area. The first one lies on

the streams and is centered on the current position along the manifold (Fig. 24 left plot, yellow dots). This is used to estimate c^{Stream} the count of stars belonging to the stream. However, the resulting count is biased by the sky contribution. In order to remove this bias we define two other apertures (Fig. 24 left plot, red dots) and estimate their respective counts as c_1^{Sky} and c_2^{Sky} . The average sky count is obtained by $\langle c^{Sky} \rangle = 0.5 \times (c_1^{Sky} + c_2^{Sky})$. The counts per pixel are obtained by normalizing the counts by the area of the Sky aperture A (fixed for all centers). We now compute the Signal of the stream by subtracting the average sky count from the stream count and obtain the SNR for the stream

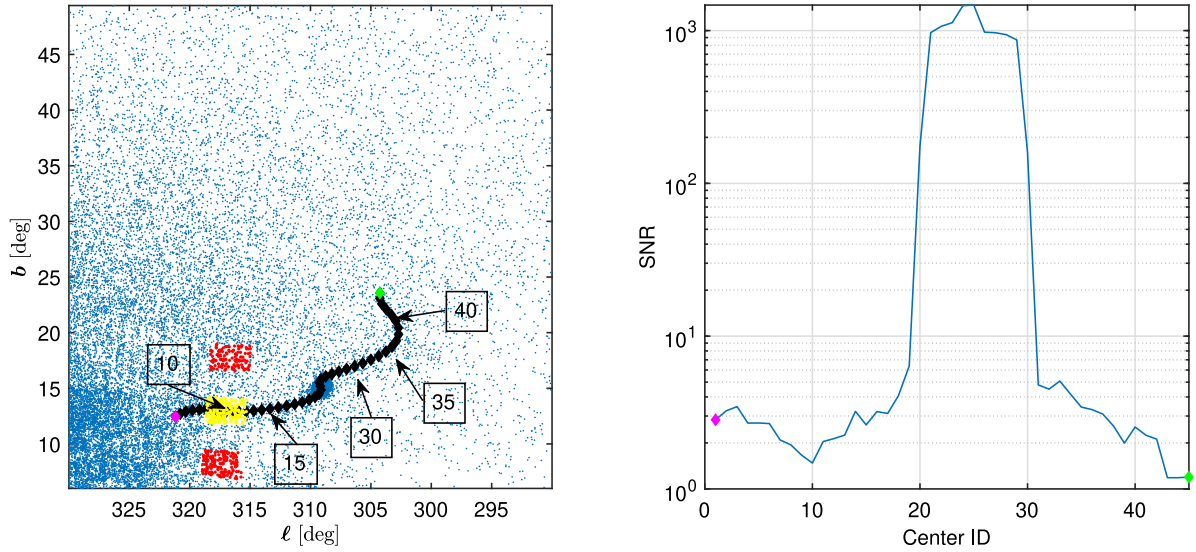


Fig. 24. Left panel: sky background selection (red circle), overlaid to the box containing the filament from ω_{Cen} . Black diamonds show the position of the centers for SGTM (Center ID), after training. Right panel: SNR for each corresponding center of SGTM. We removed the central centers identifying the core of ω_{Cen} , since our interest is on the quality of detection of the stream.

detection at point $\tilde{\mathbf{t}}_\ell \in \bar{\mathcal{P}}$:

$$\text{SNR}(\tilde{\mathbf{t}}_\ell) = \frac{S(\tilde{\mathbf{t}}_\ell)}{N(\tilde{\mathbf{t}}_\ell)} = \frac{c^{\text{Stream}} - \langle c^{\text{Sky}} \rangle \sqrt{A}}{\sqrt{\langle c^{\text{Sky}} \rangle} A}. \quad (68)$$

Particularly important for the estimation of the SNR is the area of both sky and on-object apertures. In this case, we fix the width of the rectangles to be double the distance between neighboring centers on the stream. However, for future analysis, the multiplicative factor can be changed and imposed by the user in order to optimize the result. We believe a factor 2 is sufficient in our case to obtain a reasonable SNR throughout the whole stream. The SNR for each point (Center ID in the figure) on the stream is shown in Fig. 24, right plot. In order to allow visibility of both the SNR of the streams and the core of ω -Centauri, we present the SNR in logarithmic scale. Representative Centers IDs are also shown on the left plot for an easier visual correspondence with the x-axis of the right plot. The SNR shows a sharp peak at around Center ID = 25 in correspondence to the core of ω -Centauri. The left-most edge of the plot also shows a slight increase in SNR, due to the proximity to the hosting galaxy (Center ID = 0, purple diamond). After the peak in the core, the SNR gradually decreases up to the right-most edge of the manifold (green diamond), where the filament is barely detectable.

6. Conclusions

We present a coherent, semi-automatic toolbox for the analysis of noisy data sets with underlying complicated one-dimensional structures convoluted with transverse noise. The toolbox is built on five methodologies, each one addressing different challenges in this kind of data sets. The first methodology (LAAT) aims at recovering high-density regions distributed along elongated filaments within sparse background noise. It is based on ant-colony optimization techniques and via the assignment of a scalar field over the data set it enables selection of relevant features depending on a user-specified threshold. The second methodology (EM3 A) enhances over-density along filaments, by pushing points towards the perpendicular complement of the manifolds. It is again inspired by ant-colony optimization and uses principles of game theory for parameter tuning. The third methodology (Dimensionality Index) is devoted to defining a

dimensionality index to all particles in the data set. Through the dimensionality index, it is possible to define partitions of the data set containing only points belonging to structures of defined dimensionality. A smoothed version of the index is proposed as a way to take into account the global structure to which particles locally belong. Having separated one-dimensional points from the rest of the data set, a manifold crawling algorithm is proposed that traces the skeletons of the hidden structures. In order to describe the transverse noise to the manifolds, a constrained Gaussian Mixture Model is devised in the form of Stream Generative Topographic Mapping.

We also presented two visualization techniques that take full advantage of the manifold formulation. The Bi-dimensional profile gives a global view over the manifold, showing concisely the behavior of desired quantities along the mean curve and across the radial direction of the filaments. The orthonormal coordinate frame technique gives a detailed depiction of the same quantities over local frames perpendicular to the manifold's tangent direction at each desired location.

The aim of this work is to demonstrate how the various methodologies can be combined in different ways for a range of astronomical applications, both on simulated and observed data sets. Particular care is dedicated to describing in detail the various aspects of each methodology, with the user experience in mind.

Initially, the toolbox has been applied to a mock data set, where two main filamentary structures have been defined having non-linear shapes. The point cloud sampling the manifolds includes additional transverse and background noise. Each particle in the data set has a specific value for two simulated physical quantities. The quantities are designed to follow specific profiles along and across the two filaments. The application of the toolbox to this case is used as evaluation of its accuracy in recovering the hidden structures as well as the variation of the two simulated quantities along them. Knowing the ground-truth of the two filaments, a quantitative comparison with the structures recovered via the toolbox was possible, univocally proving the quality of its performance. We also showed how removing the normalization by the “approximation to unity” from SPH interpolations, is prone to producing artifacts in the final visualization of simulated quantities, misguiding inference from simulations.

The toolbox has also been applied to two simulated data sets: a dwarf galaxy interacting with its host galaxy cluster and a filament from the cosmic web. In the first case, particular attention

has been devoted to the onset of Star Formation in the arms generated by mixing of the dwarf's interstellar gas and the gas from the cluster. After the recovery of a dominant filament in the data set, its density, neutral fraction, metallicity and temperature have been analyzed with both our visualization techniques, finding favorable Star Formation conditions in the inner parts of the manifolds, along its whole elongation. In the second application, we studied the dynamics of a filament extracted from the simulated Cosmic Web. Focusing only on Dark Matter, we show that islands of opposite curl may appear in the core of the filament, confirming previous findings. Given the orthonormal coordinate plane visualization technique, we are able to estimate the number of these islands while scrolling along the manifold. Furthermore, to prove the wide range of possible applications of our toolbox, we studied the stellar filament of the ω -Centauri globular cluster, recovered from the GAIA DR2 data set. The application to this particular data set proved successful, enabling a detailed study of the stellar number density along the manifold allowing for the computation of the local Signal-to-Noise Ratio, for a further constraint on its detectability.

For future continuation of this work, we showcase the applicability and fitness of the proposed toolbox for extensively studying the properties of the cosmic web structures. It is then possible to explore the physical properties of the Dark Matter halos, gas, or individual galaxies in relation to these structures. As also mentioned in the introduction, the ability of these methodologies to handle very large point clouds opens the possibility of investigating unexplored regions of the Milky Way that are dense in stars in pursuit of signs of merger debris. Aside from the examples mentioned so far, the proposed toolbox can be applied to a variety of cases and will help in the detection and analysis of filamentary structures with very different natures, hidden within noisy environments. It is essential to note that this also includes non-physical manifolds, (e.g. in high-dimensional parameter spaces). This work showcases the potential of 1-DREAM on manifolds in physical spaces, either simulated or observed. However, with no additional modifications or assumptions, it is possible to extend it to a variety of different spaces, such as the multidimensional parameter space of astronomical observations, or the color-magnitude diagram of stellar systems and many more. Some of these studies will be the subject of future works, but we believe that 1-DREAM will become a useful tool in many different fields.

CRediT authorship contribution statement

M. Canducci: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **P. Awad:** Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization, Investigation. **A. Taghribi:** Methodology, Software, Writing – original draft, Writing – review & editing. **M. Mohammadi:** Methodology, Software, Writing – original draft, Writing – review & editing. **M. Mastropietro:** Data curation, Software, Writing – original draft, Writing – review & editing. **S. De Rijcke:** Term, Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project acquisition, Funding acquisition. **R. Peletier:** Term, Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project acquisition, Funding acquisition. **R. Smith:** Term, Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project acquisition, Funding acquisition, Data curation. **K. Bunte:** Term, Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project acquisition, Funding acquisition. **P. Tiño:** Term, Conceptualization, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project acquisition, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721463 to the SUNDIAL ITN Network. Additional funding was also provided by the Alan Turing Institute, within the Fellowship 96102.

References

- Abbott, T.M.C., Abdalla, F.B., Allam, S., Amara, A., Annis, J., Asorey, J., Avila, S., Ballester, O., Banerji, M., Barkhouse, W., Baruah, L., Baumer, M., Bechtol, K., Becker, M.R., Benoit-Lévy, A., Bernstein, G.M., Bertin, E., Blazek, J., Bocquet, S., Brooks, D., Brout, D., Buckley-Geer, E., Burke, D.L., Busti, V., Campisano, R., Cardiel-Sas, L., Carnero Rosell, A., Carrasco Kind, M., Carretero, J., Castander, F.J., Cawthon, R., Chang, C., Chen, X., Conselice, C., Costa, G., Croce, M., Cunha, C.E., D'Andrea, C.B., da Costa, L.N., Das, R., Daues, G., Davis, T.M., Davis, C., De Vicente, J., DePoy, D.L., DeRose, J., Desai, S., Diehl, H.T., Dietrich, J.P., Dodelson, S., Doel, P., Drlica-Wagner, A., Eifler, T.F., Elliott, A.E., Evrard, A.E., Farahi, A., Fausti Neto, A., Fernandez, E., Finley, D.A., Flaugher, B., Foley, R.J., Fosalba, P., Friedel, D.N., Frieman, J., García-Bellido, J., Gaztanaga, E., Gerdes, D.W., Giannantonio, T., Gill, M.S.S., Glazebrook, K., Goldstein, D.A., Gower, M., Gruen, D., Gruendl, R.A., Gschwend, J., Gupta, R.R., Gutierrez, G., Hamilton, S., Hartley, W.G., Hinton, S.R., Hislop, J.M., Holmwood, D., Honscheid, K., Hoyle, B., Huterer, D., Jain, B., James, D.J., Jeltema, T., Johnson, M.W.G., Johnson, M.D., Kacprzak, T., Kent, S., Khullar, G., Klein, M., Kovacs, A., Koziol, A.M.G., Krause, E., Kremin, A., Kron, R., Kuehn, K., Kuhlmann, S., Kuropatkin, N., Lahav, O., Lasker, J., Li, T.S., Li, R.T., Liddle, A.R., Lima, M., Lin, H., López-Reyes, P., MacCrann, N., Maia, M.A.G., Maloney, J.D., Manera, M., March, M., Marriner, J., Marshall, J.L., Martini, P., McClintock, T., McKay, T., McMahon, R.G., Melchior, P., Menanteau, F., Miller, C.J., Miquel, R., Mohr, J.J., Morganson, E., Mould, J., Neilsen, E., Nichol, R.C., Nogueira, F., Nord, B., Nugent, P., Nunes, L., Ogando, R.L.C., Old, L., Pace, A.B., Palmese, A., Paz-Chinchón, F., Peiris, H.V., Percival, W.J., Petravick, D., Plazas, A.A., Poh, J., Pond, C., Porredon, A., Pujol, A., Refregier, A., Reil, K., Ricker, P.M., Rollins, R.P., Romer, A.K., Roodman, A., Rooney, P., Ross, A.J., Rykoff, E.S., Sako, M., Sanchez, M.L., Sanchez, E., Santiago, B., Saro, A., Scarpine, V., Scolnic, D., Serrano, S., Sevilla-Noarbe, I., Sheldon, E., Shipp, N., Silveira, M.L., Smith, M., Smith, R.C., Smith, J.A., Soares-Santos, M., Sobreira, F., Song, J., Stebbins, A., Suchyta, E., Sullivan, M., Swanson, M.E.C., Tarle, G., Thaler, J., Thomas, D., Thomas, R.C., Troxel, M.A., Tucker, D.L., Vikram, V., Vivas, A.K., Walker, A.R., Wechsler, R.H., Weller, J., Wester, W., Wolf, R.C., Wu, H., Yanny, B., Zenteno, A., Zhang, Y., Zuntz, J., DES Collaboration, Juneau, S., Fitzpatrick, M., Nikutta, R., Nidever, D., Olsen, K., Scott, A., NOAO Data Lab, 2018. The dark energy survey: Data release 1. *Astrophys. J. Suppl.* 239 (2), 18. doi:10.3847/1538-4365/aae9f0, arXiv:1801.03181.
- Allegra, M., Facco, E., Denti, F., Laio, A., Mira, A., 2020. Data segmentation based on the local intrinsic dimension. *Sci. Rep.* 10, 16449. doi:10.1038/s41598-020-72222-0, arXiv:1902.10459.
- Alpaslan, M., Robotham, A.S.G., Driver, S., Norberg, P., Baldry, I., Bauer, A.E., Bland-Hawthorn, J., Brown, M., Cluver, M., Colless, M., Foster, C., Hopkins, A., Van Kampen, E., Kelvin, L., Lara-Lopez, M.A., Liske, J., Lopez-Sanchez, A.R., Loveday, J., McNaught-Roberts, T., Merson, A., Pimblett, K., 2014. Galaxy and mass assembly (GAMA): the large-scale structure of galaxies and comparison to mock universes. *Mon. Not. R. Astron. Soc.* 438 (1), 177–194. doi:10.1093/mnras/stt2136, arXiv:1311.1211.
- Aragón-Calvo, M.A., Jones, B.J.T., van de Weygaert, R., van der Hulst, J.M., 2007. The multiscale morphology filter: identifying and extracting spatial patterns in the galaxy distribution. *Astron. Astrophys.* 474 (1), 315–338. doi:10.1051/0004-6361:20077880, arXiv:0705.2072.
- Arifanto, M.I., Fuchs, B., 2006. Fine structure in the phase space distribution of nearby subdwarfs. *Astron. Astrophys.* 449 (2), 533–538. doi:10.1051/0004-6361:20054355, arXiv:astro-ph/0512296.
- Balbinot, E., Santiago, B.X., da Costa, L.N., Makler, M., Maia, M.A.G., 2011. The tidal tails of NGC 2298. *Mon. Not. R. Astron. Soc.* 416 (1), 393–402. doi:10.1111/j.1365-2966.2011.19044.x, arXiv:1105.1933.
- Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press, pp. 585–591.

- Beygu, B., Kreckel, K., van der Hulst, T., Peletier, R., Jarrett, T., van de Weygaert, R., van Gorkom, J.H., Aragón-Calvo, M., 2016. The void galaxy survey: Morphology and star formation properties of void galaxies. In: van de Weygaert, R., Shandarin, S., Saar, E., Einasto, J. (Eds.), *The Zeldovich Universe: Genesis and Growth of the Cosmic Web*. Vol. 308. pp. 600–605. doi:10.1017/S1743921316010656, arXiv:1501.02577.
- Bishop, C.M., Svensén, M., Williams, C.K.I., 1998a. Developments of the generative topographic mapping. *Neurocomputing* 21, 203–224.
- Bishop, C.M., Svensén, M., Williams, C.K.I., 1998b. GTM: The generative topographic mapping. *Neural Comput.* 10, 215–234.
- Boissonnat, J.-D., Chazal, F., Yvinec, M., 2018. *Geometric and Topological Inference*. Cambridge University Press, Cambridge Texts in Applied Mathematics. URL <https://hal.inria.fr/hal-01615863>.
- Bonaca, A., Hogg, D.W., 2018. The information content in cold stellar streams. *Agron. J.* 867 (2), 101. doi:10.3847/1538-4357/aae4da, arXiv:1804.06854.
- Bond, J.R., Kofman, L., Pogosyan, D., 1996. How filaments of galaxies are woven into the cosmic web. *Nature* 380 (6575), 603–606. doi:10.1038/380603a0, arXiv:astro-ph/9512141.
- Bos, E.G.P., van de Weygaert, R., Dolag, K., Pettorino, V., 2012. The darkness that shaped the void: dark energy and cosmic voids. *Mon. Not. R. Astron. Soc.* 426 (1), 440–461. doi:10.1111/j.1365-2966.2012.21478.x, arXiv:1205.4238.
- Boselli, A., Gavazzi, G., 2006. Environmental effects on late-type galaxies in nearby clusters. *Publ. ASP* 118 (842), 517–559. doi:10.1086/500691, arXiv:astro-ph/0601108.
- Canducci, M., Tiño, P., Mastropietro, M., 2022. Probabilistic modelling of general noisy multi-manifold data sets. *Artificial Intelligence* 302, 103579. doi:10.1016/j.artint.2021.103579.
- Carlberg, R.G., 2012. Dark matter sub-halo counts via star stream crossings. *Agron. J.* 748 (1), 20. doi:10.1088/0004-637X/748/1/20, arXiv:1109.6022.
- Cautun, M., van de Weygaert, R., Jones, B.J.T., 2013. NEXUS: tracing the cosmic web connection. *Mon. Not. R. Astron. Soc.* 429 (2), 1286–1308. doi:10.1093/mnras/sts416, arXiv:1209.2043.
- Cautun, M., van de Weygaert, R., Jones, B.J.T., Frenk, C.S., 2014. Evolution of the cosmic web. *Mon. Not. R. Astron. Soc.* 441 (4), 2923–2973. doi:10.1093/mnras/stu768, arXiv:1401.7866.
- Chambers, K.C., Pan-STARRS Team, 2016. The pan-STARRS surveys. In: *American Astronomical Society Meeting Abstracts* #227. Vol. 227. p. 324.07.
- Chen, Y.-C., Ho, S., Mandelbaum, R., Bahcall, N.A., Brownstein, J.R., Freeman, P.E., Genovese, C.R., Schneider, D.P., Wasserman, L., 2017. Detecting effects of filaments on galaxy properties in the sloan digital sky survey III. *Mon. Not. R. Astron. Soc.* 466 (2), 1880–1893. doi:10.1093/mnras/stw3127, arXiv:1509.06376.
- Chen, H., Jiang, G., Yoshihira, K., 2006. Robust nonlinear dimensionality reduction for manifold learning. In: 18th International Conference on Pattern Recognition. Vol. 2. ICPR'06, pp. 447–450. doi:10.1109/ICPR.2006.1011.
- Chu, S.-C., Roddick, J.F., Su, C.-J., Pan, J.-S., 2004. Constrained ant colony optimization for data clustering. In: *Proceedings of the 8th Pacific Rim International Conference on Trends in Artificial Intelligence*. PRICAI '04, Springer-Verlag, Berlin, Heidelberg, pp. 534–543. doi:10.1007/978-3-540-28633-2_57.
- Cossins, P.J., 2010. *Smoothed Particle Hydrodynamics* (Ph.D. thesis). University of Leicester.
- Donoho, D.L., Grimes, C., 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci.* 100 (10), 5591–5596. doi:10.1073/pnas.1031596100.
- Dorigo, M., Di Caro, G., Gambardella, L.M., 1999. Ant algorithms for discrete optimization. *artificial life* 5, 137–172. *Artif. Life* 5, 137–172. doi:10.1162/106454699568728.
- Dorigo, M., Stützle, T., 2004. *Ant Colony Optimization*. MIT Press.
- Doroshkevich, A.G., 1980. Fragmentation of a primordial flat layer, and the formation of internal cluster structure. *Astron. Zh.* 57, 259–267.
- Duffau, S.V., Zinn, R., Carraro, G., Méndez, R.A., Vivas, A.K., Gallart, C., Winnick, R., 2006. Confirmation of halo substructure using quest RR lyrae data: The new virgo stellar stream (VSS). In: *Revista Mexicana de Astronomía Y Astrofísica Conference Series*. Vol. 26. pp. 70–71.
- Elhamifar, E., Vidal, R., 2011. Sparse manifold clustering and embedding. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc. pp. 55–63, URL <http://papers.nips.cc/paper/4246-sparse-manifold-clustering-and-embedding.pdf>.
- Facco, E., d'Errico, M., Rodriguez, A., Laio, A., 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* 7.
- Falck, B.L., Neyrinck, M.C., Szalay, A.S., 2012. ORIGAMI: Delineating halos using phase-space folds. *Agron. J.* 754 (2), 126. doi:10.1088/0004-637X/754/2/126, arXiv:1201.2353.
- Fan, M., Zhang, X., Qiao, H., Zhang, B., 2016. Efficient isometric multi-manifold learning based on the self-organizing method. *Inform. Sci.* 345 (C), 325–339. doi:10.1016/j.ins.2016.01.069.
- Fukugita, M., 1998. The sloan digital sky survey. In: *Highlights of Astronomy*. Vol. 11A. p. 449.
- Gaia Collaboration, 2018. *VizieR online data catalog: Gaia DR2* (gaia collaboration, 2018). In: *VizieR Online Data Catalog*. p. I/345.
- Gaia Collaboration, 2020. *VizieR online data catalog: Gaia EDR3* (gaia collaboration, 2020). In: *VizieR Online Data Catalog*. p. I/350.
- Gambardella, L., Dorigo, M., 1996. Solving symmetric and asymmetric TSPs by ant colonies. In: *Proceedings of IEEE International Conference on Evolutionary Computation*. pp. 622–627. doi:10.1109/ICEC.1996.542672.
- Ganeshiah Veena, P., Cautun, M., van de Weygaert, R., Tempel, E., Jones, B.J.T., Rieder, S., Frenk, C.S., 2018. The Cosmic Ballet: spin and shape alignments of haloes in the cosmic web. *Mon. Not. R. Astron. Soc.* 481 (1), 414–438. doi:10.1093/mnras/sty2270, arXiv:1805.00033.
- Gingold, R., Monaghan, J., 1977. Smoothed particle hydrodynamics - theory and application to non-spherical stars. *Mon. Not. R. Astron. Soc.* 181, 375–389. doi:10.1093/mnras/181.3.375.
- Grossi, M., 2018. Evolution of star-forming dwarf galaxies in different environments. In: *Proceedings of the International Astronomical Union*. Vol. 14. (S344), Cambridge University Press, pp. 319–330. doi:10.1017/S1743921318007159.
- Gunn, J.E., Carr, M., Rockosi, C., Sekiguchi, M., Berry, K., Elms, B., de Haas, E., Ivezić, Ž., Knapp, G., Lupton, R., Pauls, G., Simcoe, R., Hirsch, R., Sanford, D., Wang, S., York, D., Harris, F., Annis, J., Bartoček, L., Boroski, W., Bakken, J., Haldeman, M., Kent, S., Holm, S., Holmgren, D., Petrávick, D., Prosapio, A., Rechenmacher, R., Doi, M., Fukugita, M., Shimasaku, K., Okada, N., Hull, C., Siegmund, W., Mannery, E., Blouke, M., Heidtman, D., Schneider, D., Lucinio, R., Brinkman, J., 1998. The sloan digital sky survey photometric camera. *Astron. J.* 116 (6), 3040–3081. doi:10.1086/300645, arXiv:astro-ph/9809085.
- Hahn, O.J., 2009. *Galaxy Formation in the Cosmic Web* (Ph.D. thesis). ETH Zurich, Switzerland.
- Hahn, O., Abel, T., 2011. Multi-scale initial conditions for cosmological simulations. *Mon. Not. R. Astron. Soc.* 415 (3), 2101–2121. doi:10.1111/j.1365-2966.2011.18820.x, arXiv:1103.6031.
- Hahn, O., Carollo, C.M., Porciani, C., Dekel, A., 2007a. The evolution of dark matter halo properties in clusters, filaments, sheets and voids. *Mon. Not. R. Astron. Soc.* 381 (1), 41–51. doi:10.1111/j.1365-2966.2007.12249.x, arXiv:0704.2595.
- Hahn, O., Porciani, C., Carollo, C.M., Dekel, A., 2007b. Properties of dark matter haloes in clusters, filaments, sheets and voids. *Mon. Not. R. Astron. Soc.* 375 (2), 489–499. doi:10.1111/j.1365-2966.2006.11318.x, arXiv:astro-ph/0610280.
- Hamilton, W.R., 1866. *Elements of Quaternions*. Longmans, Green, & Company.
- Han, X.-F., Jin, J.S., Wang, M.-J., Jiang, W., Gao, L., Xiao, L., 2017. A review of algorithms for filtering the 3D point cloud. *Signal Process., Image Commun.* 57, 103–112. doi:10.1016/j.image.2017.05.009, URL <https://www.sciencedirect.com/science/article/pii/S0923596517300930>.
- Hao, Z., Liu, J., Ma, S., Jin, X., Lian, X., 2017. Noise-removal method for manifold learning. In: Yue, D., Peng, C., Du, D., Zhang, T., Zheng, M., Han, Q. (Eds.), *Intelligent Computing, Networked Control, and their Engineering Applications*. Springer Singapore, Singapore, pp. 191–200.
- Haro, G., Randal, G., Sapiro, G., Haro, G., Randall, G., Sapiro, G., 2000. Translated Poisson mixture model for stratification learning. *Int. J. Comput. Vis.*
- Hein, M., Maier, M., 2007. Manifold denoising. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems*. Vol. 19. MIT Press, pp. 561–568, URL <https://proceedings.neurips.cc/paper/2006/file/a0b83c02d720415dada82e08bc09e9f3-Paper.pdf>.
- Helmi, A., 2020. Streams, substructures, and the early history of the milky way. *Ann. Rev. Astron. Astrophys.* 58, 205–256. doi:10.1146/annurev-astro-032620-021917, arXiv:2002.04340.
- Hettiarachchi, R., Peters, J., 2015. Multi-manifold LLE learning in pattern recognition. *Pattern Recognit.* 48 (9), 2947–2960. doi:10.1016/j.patcog.2015.04.003.
- Huo, X., Ni, X.S., Smith, A.K., 2008. A survey of manifold-based learning methods. In: *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*. pp. 691–745. doi:10.1142/9789812779861_0015.
- Ibata, R.A., Bellazzini, M., Malhan, K., Martin, N., Bianchini, P., 2019a. Identification of the long stellar stream of the prototypical massive globular cluster ω Centauri. *Nat. Astron.* 3, 667–672. doi:10.1038/s41550-019-0751-x, arXiv:1902.09544.
- Ibata, R., Irwin, M., Lewis, G., Ferguson, A.M.N., Tanvir, N., 2001. A giant stream of metal-rich stars in the halo of the galaxy M31. *Nature* 412, 49–52. doi:10.1038/35083506, arXiv:astro-ph/0107090.
- Ibata, R.A., Lewis, G.F., Irwin, M.J., Quinn, T., 2002. Uncovering cold dark matter halo substructure with tidal streams. *Mon. Notices RAS* 332 (4), 915–920. doi:10.1046/j.1365-8711.2002.05358.x, arXiv:astro-ph/0110690.
- Ibata, R.A., Malhan, K., Martin, N.F., 2019b. The streams of the gaping abyss: A population of entangled stellar streams surrounding the inner galaxy. *Agron. J.* 872 (2), 152. doi:10.3847/1538-4357/ab0080, arXiv:1901.07566.
- Ibata, R., Malhan, K., Martin, N., Aubert, D., Famaey, B., Bianchini, P., Monari, G., Siebert, A., Thomas, G.F., Bellazzini, M., Bonifacio, P., Caffau, E., Renaud, F., 2021. Charting the galactic acceleration field. I. A search for stellar streams with Gaia DR2 and EDR3 with follow-up from ESPaDOnS and UVES. *Agron. J.* 914 (2), 123. doi:10.3847/1538-4357/abfcc2, arXiv:2012.05245.

- Johnston, K.V., 1999. The role of accretion in forming the galactic halo. In: Gibson, B.K., Axelrod, R.S., Putman, M.E. (Eds.), *The Third Stromlo Symposium: The Galactic Halo*. In: *Astronomical Society of the Pacific Conference Series*, vol. 165, p. 64. doi:10.48550/ARXIV.ASTRO-PH/9810421, arXiv:astro-ph/9810421.
- Johnston, K.V., Hernquist, L., Bolte, M., 1996. Fossil signatures of ancient accretion events in the halo. *Agron. J.* 465, 278. doi:10.1086/177418, arXiv:astro-ph/9602060.
- Johnston, K.V., Spergel, D.N., Haydn, C., 2002. How lumpy is the milky way's dark matter halo? *Agron. J.* 570 (2), 656–664. doi:10.1086/339791, arXiv:astro-ph/0111196.
- Jones, B.J.T., van de Weygaert, R., Aragón-Calvo, M.A., 2010. Fossil evidence for spin alignment of sloan digital sky survey galaxies in filaments. *Mon. Not. R. Astron. Soc.* 408 (2), 897–918. doi:10.1111/j.1365-2966.2010.17202.x, arXiv:1001.4479.
- Kaslovsky, D.N., Meyer, F.G., 2014. Non-asymptotic analysis of tangent space perturbation. *Inform. Inference A J. IMA* 3 (2), 134–187. doi:10.1093/imaia/iau004.
- Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43 (1), 59–69.
- Kraljic, K., Arnouts, S., Pichon, C., Laigle, C., de la Torre, S., Vibert, D., Cadiou, C., Dubois, Y., Treyer, M., Schimd, C., Codis, S., de Lapparent, V., Devriendt, J., Hwang, H.S., Le Borgne, D., Malavasi, N., Milliard, B., Musso, M., Pogossyan, D., Alpaslan, M., Bland-Hawthorn, J., Wright, A.H., 2018. Galaxy evolution in the metric of the cosmic web. *Mon. Not. R. Astron. Soc.* 474 (1), 547–571. doi:10.1093/mnras/stx2638, arXiv:1710.02676.
- Kupper, A.H.W., Balbinot, E., Bonaca, A., Johnston, K.V., Hogg, D., Kroupa, P., Santiago, B., 2015. Globular cluster streams as galactic high-precision scales. In: *IAU General Assembly*, Vol. 29. 2258435.
- Laigle, C., Pichon, C., Codis, S., Dubois, Y., Le Borgne, D., Pogossyan, D., Devriendt, J., Peirani, S., Prunet, S., Rouberol, S., Slyz, A., Sousbie, T., 2014. Swirling around filaments: are large-scale structure vortices spinning up dark haloes? *Mon. Not. R. Astron. Soc.* 446 (3), 2744–2759. doi:10.1093/mnras/stu2289.
- Lavaux, G., Wandelt, B.D., 2010. Precision cosmology with voids: definition, methods, dynamics. *Mon. Not. R. Astron. Soc.* 403 (3), 1392–1408. doi:10.1111/j.1365-2966.2010.16197.x, arXiv:0906.4101.
- Lee, J., Park, D., 2009. Constraining the dark energy equation of state with cosmic voids. *Astrophys. J.* 696 (1), L10–L12. doi:10.1088/0004-637X/696/1/L10, arXiv:0704.0881.
- Lewis, A., Challinor, A., 2011. CAMB: Code for anisotropies in the microwave background. *ascl*:1102.026. arXiv:1102.026.
- Libeskind, N.I., van de Weygaert, R., Cautun, M., Falck, B., Tempel, E., Abel, T., Alpaslan, M., Aragón-Calvo, M.A., Forero-Romero, J.E., Gonzalez, R., Gottlöber, S., Hahn, O., Hellwing, W.A., Hoffman, Y., Jones, B.J.T., Kitaura, F., Knebe, A., Mantzi, S., Neyrinck, M., Nuza, S.E., Padilla, N., Platen, E., Ramachandra, N., Robotham, A., Saar, E., Shandarin, S., Steinmetz, M., Stoica, R.S., Sousbie, T., Yepes, G., 2018. Tracing the cosmic web. *Mon. Not. R. Astron. Soc.* 473 (1), 1195–1217. doi:10.1093/mnras/stx1976, arXiv:1705.03021.
- Lin, T., Zha, H., 2008. Riemannian manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 30 (5), 796–809. doi:10.1109/TPAMI.2007.70735.
- Lyu, H., Sha, N., Qin, S., Yan, M., Xie, Y., Wang, R., 2019. Manifold denoising by nonlinear robust principal component analysis. doi:10.48550/ARXIV.1911.03831, arXiv preprint arXiv:1911.03831.
- Majewski, S.R., 1999. The role of accretion in the formation of the halo: Observational view. In: Gibson, B.K., Axelrod, R.S., Putman, M.E. (Eds.), *The Third Stromlo Symposium: The Galactic Halo*. In: *Astronomical Society of the Pacific Conference Series*, vol. 165, p. 76.
- Malhan, K., Ibata, R.A., 2018. STREAMFINDER - I. A new algorithm for detecting stellar streams. *Mon. Not. R. Astron. Soc.* 477 (3), 4063–4076. doi:10.1093/mnras/sty912, arXiv:1804.11338.
- Malhan, K., Ibata, R.A., 2019. Constraining the Milky way halo potential with the GD-1 stellar stream. *Mon. Not. R. Astron. Soc.* 486 (3), 2995–3005. doi:10.1093/mnras/stz1035, arXiv:1807.05994.
- Martínez-Delgado, D., Gabany, R.J., Crawford, K., Zibetti, S., Majewski, S.R., Rix, H.-W., Fliri, J., Carballo-Bello, J.A., Bardalez-Gagliuffi, D.C., Peñarrubia, J., Chonis, T.S., Madore, B., Trujillo, I., Schirmer, M., McDavid, D.A., 2010. Stellar tidal streams in spiral galaxies of the local volume: A pilot survey with modest aperture telescopes. *Astron. J.* 140 (4), 962–967. doi:10.1088/0004-6256/140/4/962, arXiv:1003.4860.
- Mastropietro, M., De Rijcke, S., Peletier, R.F., 2021. A tale of two tails: insights from simulations into the formation of the peculiar dwarf galaxy NGC 1427A. *Mon. Not. R. Astron. Soc.* 504 (3), 3387–3398. doi:10.1093/mnras/stab1091, arXiv:2104.07671.
- Mayer, L., Mastropietro, C., Wadsley, J., Stadel, J., Moore, B., 2006. Simultaneous ram pressure and tidal stripping: how dwarf spheroidals lost their gas. *Mon. Not. R. Astron. Soc.* 369 (3), 1021–1038. doi:10.1111/j.1365-2966.2006.10403.x, arXiv:astro-ph/0504277.
- McConnachie, A.W., Irwin, M.J., Ibata, R.A., Dubinski, J., Widrow, L.M., Martin, N.F., Côté, P., Dotter, A.L., Navarro, J.F., Ferguson, A.M.N., Puzia, T.H., Lewis, G.F., Babul, A., Barmby, P., Bienaymé, O., Chapman, S.C., Cockcroft, R., Collins, M.L.M., Fardal, M.A., Harris, W.E., Huxor, A., Mackey, A.D., Peñarrubia, J., Rich, R.M., Richer, H.B., Siebert, A., Tanvir, N., Valls-Gabaud, D., Venn, K.A., 2009. The remnants of galaxy formation from a panoramic survey of the region around M31. *Nature* 461 (7260), 66–69. doi:10.1038/nature08327, arXiv:0909.0398.
- McPartland, C., Ebeling, H., Roediger, E., Blumenthal, K., 2016. Jellyfish: the origin and distribution of extreme ram-pressure stripping events in massive galaxy clusters. *Mon. Not. R. Astron. Soc.* 455 (3), 2994–3008. doi:10.1093/mnras/stv2508, arXiv:1511.00033.
- Mohammadi, M., Bunte, K., 2020. Multi-agent based manifold denoising. In: Analide, C., Novais, P., Camacho, D., Yin, H. (Eds.), *Intelligent Data Engineering and Automated Learning - IDEAL 2020*. Springer International Publishing, Cham, pp. 12–24.
- Mohammadi, M., Tino, P., Bunte, K., 2021. Manifold alignment aware ants: a Markovian process for manifold extraction. *Neural Comput.*
- Mordohai, P., Medioni, G., 2005. Unsupervised dimensionality estimation and manifold learning in high-dimensional spaces by tensor voting. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence. IJCAI '05*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 798–803.
- Mordohai, P., Medioni, G., 2010. Dimensionality estimation, manifold learning and function approximation using tensor voting. *J. Mach. Learn. Res.* 11, 411–450.
- Mori, M., Burkert, A., 2000. Gas stripping of dwarf galaxies in clusters of galaxies. *Agron. J.* 538 (2), 559–568. doi:10.1086/309140, arXiv:astro-ph/0001422.
- Park, D., Lee, J., 2007. Void ellipticity distribution as a probe of cosmology. *Phys. Rev. Lett.* 98 (8), 081301. doi:10.1103/PhysRevLett.98.081301, arXiv:astro-ph/0610520.
- Park, J., Zhang, Z., Zha, H., Kasturi, R., 2004. Local smoothing for manifold learning. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. Vol. 2. p. II. doi:10.1109/CVPR.2004.1315199.
- Pauls, J.L., Melott, A.L., 1995. Hierarchical pancaking: why the zel'dovich approximation describes coherent large-scale structure in N-body simulations of gravitational clustering. *Mon. Not. R. Astron. Soc.* 274 (1), 99–109. doi:10.1093/mnras/274.1.99, arXiv:astro-ph/9408019.
- Paz, D.J., Stasyszyn, F., Padilla, N.D., 2008. Angular momentum-large-scale structure alignments in Λ CDM models and the SDSS. *Mon. Not. R. Astron. Soc.* 389 (3), 1127–1136. doi:10.1111/j.1365-2966.2008.13655.x, arXiv:0804.4477.
- Pearson, S., Küpper, A.H.W., Johnston, K.V., Price-Whelan, A.M., 2015. Tidal stream morphology as an indicator of dark matter halo geometry: The case of palomar 5. *Agron. J.* 799 (1), 28. doi:10.1088/0004-637X/799/1/28, arXiv:1410.3477.
- Peebles, P.J.E., 1980. *The Large-Scale Structure of the Universe*.
- Pisani, A., Sutter, P.M., Hamaus, N., Alizadeh, E., Biswas, R., Wandelt, B.D., Hirata, C.M., 2015. Counting voids to probe dark energy. *Phys. Rev. D* 92 (8), 083531. doi:10.1103/PhysRevD.92.083531, arXiv:1503.07690.
- Platen, E., van de Weygaert, R., Jones, B.J.T., 2008. Alignment of voids in the cosmic web. *Mon. Not. R. Astron. Soc.* 387 (1), 128–136. doi:10.1111/j.1365-2966.2008.13019.x, arXiv:0711.2480.
- Price, D.J., 2012. Smoothed particle hydrodynamics and magnetohydrodynamics. *J. Comput. Phys.* 231 (3), 759–794. doi:10.1016/j.jcp.2010.12.011, Special Issue: Computational Plasma Physics.
- Rockosi, C.M., Odenkirchen, M., Grebel, E.K., Dehnen, W., Cudworth, K.M., Gunn, J.E., York, D.G., Brinkmann, J., Hennessy, G.S., Ivezić, Ž., 2002. A matched-filter analysis of the tidal tails of the globular cluster palomar 5. *Astron. J.* 124 (1), 349–363. doi:10.1086/340957.
- Roediger, E., Brüggen, M., 2008. Ram pressure stripping in a viscous intracluster medium. *Mon. Not. R. Astron. Soc.* 388 (1), L89–L93. doi:10.1111/j.1745-3933.2008.00506.x, arXiv:0806.1406.
- Roediger, E., Kraft, R.P., Nulsen, P.E.J., Forman, W.R., Machacek, M., Randall, S., Jones, C., Churazov, E., Kokotanekova, R., 2015. Stripped elliptical galaxies as probes of ICM physics: I. Tails, wakes, and flow patterns in and around stripped ellipticals. *Agron. J.* 806 (1), 103. doi:10.1088/0004-637X/806/1/103, arXiv:1409.6300.
- Rojas, R.R., Vogeley, M.S., Hoyle, F., Brinkmann, J., 2004. Photometric properties of void galaxies in the sloan digital sky survey. *Astrophys. J.* 617 (1), 50–63. doi:10.1086/425225, arXiv:astro-ph/0307274.
- Roman-Oliveira, F., Chies-Santos, A.L., Ferrari, F., Lucatelli, G., Rodríguez Del Pino, B., 2021. Morphometry as a probe of the evolution of jellyfish galaxies: evidence of broadening in the surface brightness profiles of ram-pressure stripping candidates in the multicluster system A901/A902. *Mon. Not. R. Astron. Soc.* 500 (1), 40–53. doi:10.1093/mnras/staa3226.
- Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326.
- Sathyaprakash, B.S., Sahni, V., Shandarin, S.F., 1996. Emergence of filamentary structure in cosmological gravitational clustering. *Astrophys. J. Lett.* 462, L5. doi:10.1086/310024, arXiv:astro-ph/9603085.

- Schlafly, E.F., Finkbeiner, D.P., 2011. Measuring reddening with sloan digital sky survey stellar spectra and recalibrating SFD. *Agron. J.* 737 (2), 103. doi:10.1088/0004-637X/737/2/103, arXiv:1012.4804.
- Schlegel, D.J., Finkbeiner, D.P., Davis, M., 1998. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Agron. J.* 500 (2), 525–553. doi:10.1086/305772, arXiv:astro-ph/9710327.
- Shapiro, P.R., Struck-Marcell, C., Melott, A.L., 1983. Pancakes and the formation of galaxies in a neutrino-dominated universe. *Agron. J.* 275, 413–429. doi:10.1086/161543.
- Springel, V., 2005. The cosmological simulation code gadget-2. *Mon. Not. R. Astron. Soc.* 364 (4), 1105–1134. doi:10.1111/j.1365-2966.2005.09655.x.
- Steinhauser, D., Schindler, S., Springel, V., 2016. Simulations of ram-pressure stripping in galaxy-cluster interactions. *Astron. Astrophys.* 591, A51. doi:10.1051/0004-6361/201527705, arXiv:1604.05193.
- Steyrleithner, P., Hensler, G., Boselli, A., 2020. The effect of ram-pressure stripping on dwarf galaxies. *Mon. Not. R. Astron. Soc.* 494 (1), 1114–1127. doi:10.1093/mnras/staa775, arXiv:2003.09591.
- Sutter, P.M., Carlesi, E., Wandelt, B.D., Knebe, A., 2015. On the observability of coupled dark energy with cosmic voids. *Mon. Not. R. Astron. Soc.* 446, L1–L5. doi:10.1093/mnras/slu155, arXiv:1406.0511.
- Taghribi, A., Bunte, K., Smith, R., Shin, J., Mastropietro, M., Peletier, R.F., Tino, P., 2022. LAAT: Locally aligned ant technique for discovering multiple faint low dimensional structures of varying density. *IEEE Trans. Knowl. Data Eng.* 1. doi:10.1109/TKDE.2022.3177368.
- Tempel, E., Stoica, R.S., Kipper, R., Saar, E., 2016. Bisous model-detecting filamentary patterns in point processes. *Astron. Comput.* 16, 17–25. doi:10.1016/j.ascom.2016.03.004, arXiv:1603.08957.
- Tempel, E., Stoica, R.S., Saar, E., 2013. Evidence for spin alignment of spiral and elliptical/S0 galaxies in filaments. *Mon. Not. R. Astron. Soc.* 428 (2), 1827–1836. doi:10.1093/mnras/sts162, arXiv:1207.0068.
- Tenenbaum, J.B., Silva, V.d., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290 (5500), 2319–2323. doi:10.1126/science.290.5500.2319.
- Thomas, G.F., Famaey, B., Ibata, R., Lüghausen, F., Kroupa, P., 2017. Stellar streams as gravitational experiments. I. The case of Sagittarius. *Astron. Astrophys.* 603, A65. doi:10.1051/0004-6361/201730531, arXiv:1705.01552.
- Thomas, G.F., Famaey, B., Ibata, R., Renaud, F., Martin, N.F., Kroupa, P., 2018. Stellar streams as gravitational experiments. II. Asymmetric tails of globular cluster streams. *Astron. Astrophys.* 609, A44. doi:10.1051/0004-6361/201731609, arXiv:1709.01934.
- Ting, D., Jordan, M.I., 2020. Manifold learning via manifold deflation. doi:10.48550/ARXIV.2007.03315, arXiv:2007.03315.
- Tino, P., Nabney, I., 2002. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (5), 639–656.
- Tonnesen, S., Bryan, G.L., 2012. Star formation in ram pressure stripped galactic tails. *Mon. Not. RAS* 422 (2), 1609–1624. doi:10.1111/j.1365-2966.2012.20737.x, arXiv:1203.0308.
- Tu, L., 2010. An Introduction to Manifolds. In: Universitext, Springer, New York, URL <https://books.google.co.uk/books?id=br1KngEACAAJ>.
- Vera-Casanova, A., Gómez, F.A., Monachesi, A., Gargiulo, I., Pallero, D., Grand, R.J.J., Marinacci, F., Pakmor, R., Simpson, C.M., Frenk, C.S., Morales, G., 2021. Linking the brightest stellar streams with the accretion history of Milky way-like galaxies. doi:10.1093/mnras/stac1636, arXiv e-prints. arXiv:2105.06467.
- Verbeke, R., Papastergis, E., Ponomareva, A.A., Rath, S., De Rijcke, S., 2017. A new astrophysical solution to the too big to fail problem. Insights from the moria simulations. *Astron. Astrophys.* 607, A13. doi:10.1051/0004-6361/201730758, arXiv:1703.03810.
- Verbeke, R., Vandenbroucke, B., De Rijcke, S., 2015. How the first stars shaped the faintest gas-dominated dwarf galaxies. *Astrophys. J.* 815, 85. doi:10.1088/0004-637X/815/2/85, arXiv:1511.01484.
- Wang, W., Carreira-Perpiñán, M.A., 2010. Manifold blurring mean shift algorithms for manifold denoising. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1759–1766. doi:10.1109/CVPR.2010.5539845.
- Wang, X., Tiño, P., Fardal, M.A., 2008. Multiple manifolds learning framework based on hierarchical mixture density model. In: Proceedings of the 2008th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II. ECMLPKDD '08, Springer-Verlag, Berlin, Heidelberg, pp. 566–581.
- Welker, C., Bland-Hawthorn, J., Van de Sande, J., Lagos, C., Elahi, P., Obreschkow, D., Bryant, J., Pichon, C., Cortese, L., Richards, S.N., Croom, S.M., Goodwin, M., Lawrence, J.S., Sweet, S., Lopez-Sanchez, A., Medling, A., Owers, M.S., Dubois, Y., Devriendt, J., 2019. The SAMI galaxy survey: first detection of a transition in spin orientation with respect to cosmic filaments in the stellar kinematics of galaxies. *Mon. Not. R. Astron. Soc.* 491 (2), 2864–2884. doi:10.1093/mnras/stz2860.
- Williams, M.E.K., Steinmetz, M., Sharma, S., Bland-Hawthorn, J., de Jong, R.S., Seabroke, G.M., Helmi, A., Freeman, K.C., Binney, J., Minchev, I., Bienaymé, O., Campbell, R., Fulbright, J.P., Gibson, B.K., Gilmore, G.F., Grebel, E.K., Munari, U., Navarro, J.F., Parker, Q.A., Reid, W., Siebert, A., Siviero, A., Watson, F.G., Wyse, R.F.G., Zwitter, T., 2011. The dawning of the stream of aquarius in RAVE. *Agron. J.* 728 (2), 102. doi:10.1088/0004-637X/728/2/102, arXiv:1012.2127.
- Winkel, N., Pasquali, A., Kraljic, K., Smith, R., Gallazzi, A., Jackson, T.M., 2021. The imprint of cosmic web quenching on central galaxies. *Mon. Not. R. Astron. Soc.* 505 (4), 4920–4934. doi:10.1093/mnras/stab1562, arXiv:2105.13368.
- Wu, S., Bertholet, P., Huang, H., Cohen-Or, D., Gong, M., Zwicker, M., 2018. Structure-aware data consolidation. *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (10), 2529–2537. doi:10.1109/TPAMI.2017.2754254.
- Yao, Z., Xia, Y., 2019. Manifold fitting under unbounded noise. doi:10.48550/ARXIV.1909.10228, arXiv preprint arXiv:1909.10228.
- York, D.G., Adelman, J., Anderson, J., Anderson, S.F., Annis, J., Bahcall, N.A., Bakken, J.A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W.N., Bracker, S., Briegel, C., Briggs, J.W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M.A., Castander, F.J., Chen, B., Colestock, P.L., Connolly, A.J., Crocker, J.H., Csabai, I., Czarapata, P.C., Davis, J.E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B.R., Evans, M.L., Fan, X., Federwitz, G.R., Fiscelli, L., Friedman, S., Frieman, J.A., Fukugita, M., Gillespie, B., Gunn, J.E., Gurbani, V.K., de Haas, E., Haldeman, M., Harris, F.H., Hayes, J., Heckman, T.M., Hennessey, G.S., Hindsley, R.B., Holm, S., Holmgren, D.J., Huang, C.-h., Hull, C., Husby, D., Ichikawa, S.-I., Ichikawa, T., Ivezić, Ž., Kent, S., Kim, R.S.J., Kinney, E., Klaene, M., Kleinman, A.N., Kleinman, S., Knapp, G.R., Korienek, J., Kron, R.G., Kunszt, P.Z., Lamb, D.Q., Lee, B., Leger, R.F., Limmongkol, S., Lindenmeyer, C., Long, D.C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R.H., MacKinnon, B., Mannery, E.J., Mantsch, P.M., Margon, B., McGehee, P., McKay, T.A., Meiksin, A., Merelli, A., Monet, D.G., Munn, J.A., Narayanan, V.K., Nash, T., Neilsen, E., Neswold, R., Newberg, H.J., Nichol, R.C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J.P., Owen, R., Pauls, A.G., Peoples, J., Peterson, R.L., Petravick, D., Pier, J.R., Pope, A., Pordes, R., Prossapio, A., Rechenmacher, R., Quinn, T.R., Richards, G.T., Richmond, M.W., Rivetta, C.H., Rockosi, C.M., Ruthmansdorfer, K., Sandford, D., Schlegel, D.J., Schneider, D.P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W.A., Smee, S., Smith, J.A., Snedden, S., Stone, R., Stoughton, C., Strauss, M.A., Stubbs, C., SubbaRao, M., Szalay, A.S., Szapudi, I., Szokoly, G.P., Thakar, A.R., Tremonti, C., Tucker, D.L., Uomoto, A., Vanden Berk, D., Vogeley, M.S., Waddell, P., Wang, S.-i., Watanabe, M., Weinberg, D.H., Yanny, B., Yasuda, N., SDSS Collaboration, 2000. The sloan digital sky survey: Technical summary. *Astron. J.* 120 (3), 1579–1587. doi:10.1086/301513, arXiv:astro-ph/0006396.
- Yun, K., Pillepich, A., Zinger, E., Nelson, D., Donnari, M., Joshi, G., Rodriguez-Gomez, V., Genel, S., Weinberger, R., Vogelsberger, M., Hernquist, L., 2019. Jellyfish galaxies with the IllustrisTNG simulations - I. gas-stripping phenomena in the full cosmological context. *Mon. Not. R. Astron. Soc.* 483 (1), 1042–1066. doi:10.1093/mnras/sty3156, arXiv:1810.00005.