RATS: Unsupervised manifold learning using low-distortion alignment of tangent spaces

Dhruv Kohli[1,*], Johannes S. Nieuwenhuis[2,3*], Alexander Cloninger[1,4], Gal Mishne[4#] & Devika Narain[2,3#]

Affiliations

[1] Department of Mathematics, University of California San Diego, San Diego, USA

[2] Dept. of Neuroscience, Erasmus University Medical Center, Rotterdam, The Netherlands

[3] Faculty of Applied Sciences, Technical University of Delft, Delft, The Netherlands

[4] Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, USA

[*] These authors contributed equally.

[#] These authors contributed equally.

Correspondence can be addressed to gmishne@ucsd.edu and d.narain@erasmusmc.nl.

Abstract: With the ubiquity of high-dimensional datasets in various biological fields, identifying low-dimensional topological manifolds within such datasets may reveal principles connecting latent variables to measurable instances in the world. The reliable discovery of such manifold structure in high-dimensional datasets can prove challenging, however, largely due to the introduction of distortion by leading manifold learning methods. The problem is further exacerbated by the lack of consensus on how to evaluate the quality of the recovered manifolds. Here, we present a novel measure of distortion to evaluate low-dimensional representations obtained using different techniques. We additionally develop a novel bottom-up manifold learning technique called Riemannian Alignment of Tangent Spaces (RATS) that aims to recover low-distortion embeddings of data, including the ability to embed closed manifolds into their intrinsic dimension using a unique tearing process. Compared to previous methods, we show that RATS provides low-distortion embeddings that excel in the visualization and deciphering of latent variables across a range of idealized, biological, and surrogate datasets that mimic real-world data.

One-sentence summary:

We introduce a novel dimensionality reduction technique that generates low-dimensional embeddings while preserving the global structure within the data for a variety of biological and non-biological datasets.

Introduction:

The manifold hypothesis is prevalent in many biological and non-biological fields, where complex, multifaceted datasets are thought to be well-represented by latent lower-dimensional manifolds despite being observed in high-dimensional feature spaces. Such manifold-like structures have emerged in several empirical applications within computational domains such as computer vision[1], language and representation learning[2–4], and in biological domains such as gene expression[5–8], oncology[9], botany[10], and neuroscience[11–16]. These structures are often revealed using manifold learning techniques, which aim to learn a lower-dimensional representation of data, ideally capturing the *intrinsic dimension* of the dataset. The intrinsic dimension refers to the minimum number of variables needed to effectively describe the data.

Recent research in biological and non-biological fields[17–20] has demonstrated valuable insights gained through the application of unsupervised manifold learning techniques and their derivatives, such as t-distributed Stochastic Neighbor Embedding[21] (t-SNE), Uniform Manifold Approximation and Projection[22] (UMAP), Isometric feature mapping[23] (ISOMAP), Local Tangent Space Alignment[24] (LTSA) and Low Distortion Local Eigenmaps[25] (LDLE), among others. Recent work has also adopted supervised methods for manifold learning[26]. These techniques differ in their assumptions regarding preserved geometric properties and in their choice of trade-offs while computing low-dimensional embeddings. All of these methods, however, incur significant distortion of distances along the original data manifold when mapping into the embedding space. Such distortion often arises from the uneven expansion or contraction of local spaces around embedded data points[27] (Figure 1a). This is problematic for example, when we expect the latent variables of the manifold to have physical meaning, such as space or time. In cases where the measured data does not fully represent such physical quantities, we may be unable to disambiguate such a scenario from methods that inadvertently introduce distortions, resulting in a skewed interpretation of relationships with external variables. This can have undesirable effects, such as reduced reliability in decoding how biological variables represent phenotypes and behaviors. Consequently, erroneous conclusions may be drawn due to such algorithmic artifacts. In this work, we seek to mitigate this problem in two ways. First, we present a novel measure to quantify distortion in low-dimensional embeddings across different manifold learning techniques. Second, we develop a new bottom-up manifold learning method called Riemannian Alignment of Tangent Spaces (RATS) for the unsupervised discovery of low-distortion embeddings of complex datasets.

A novel measure of distortion

Different measures have been proposed to quantify the quality of low-dimensional embeddings. Some of these utilize downstream tasks[17], for example evaluating whether cluster memberships between the ambient dimension and embedding are preserved (*e.g.*, measuring the overlap between k-nearest clusters in the embedding and data), or evaluating clustering accuracy (*e.g.*, silhouette) in the embedding space. Others evaluate the preservation of local information, for example calculating the overlap between the k-nearest neighbors in the original high-dimensional data that are preserved as k-nearest neighbors in the embedding. However, distances can still be distorted even if the ranking of neighbors is preserved. Local distortion measures[27], for example, the average or worst-case contraction and expansion of Euclidean distances of local neighborhoods in the embedding space, with respect to distances of those neighborhoods in the ambient space. However, a quantitative analysis of the global distortion incurred in these embeddings is often lacking in the literature. Here, we present a natural way to quantify the global distortion introduced when embedding data into a low-dimensional space based on the disparity between the geodesics (shortest paths between pairs of points) in the data $\{x_k\}_1^n$ and the corresponding geodesics in the low-dimensional embedding, $\{y_k\}_1^n$ (Figure 1a,b). In general, the geodesic distances in the embedding differ from their counterparts in the data due to warping induced upon embedding data into lower dimensions. We consider that any distortion between these distances arises from expansions and contractions of geodesic lengths rather than from isometric scaling. To capture such contractions and expansions, we present a measure of *Global distortion*, $\mathbb{G}_k$, defined at the kth point as the ratio of the worst expansion and the worst contraction of the shortest paths originating from x_k (Figure 1b). This measure is invariant to isometric transformations and scaling of the data, i.e. rotating the dataset or scaling all the points by the constant will result in global distortion of 1 for all points. Considering the global distortion of all points in the data generates a distribution of distortions (Figure 1b), enabling a more comprehensive evaluation of embeddings.

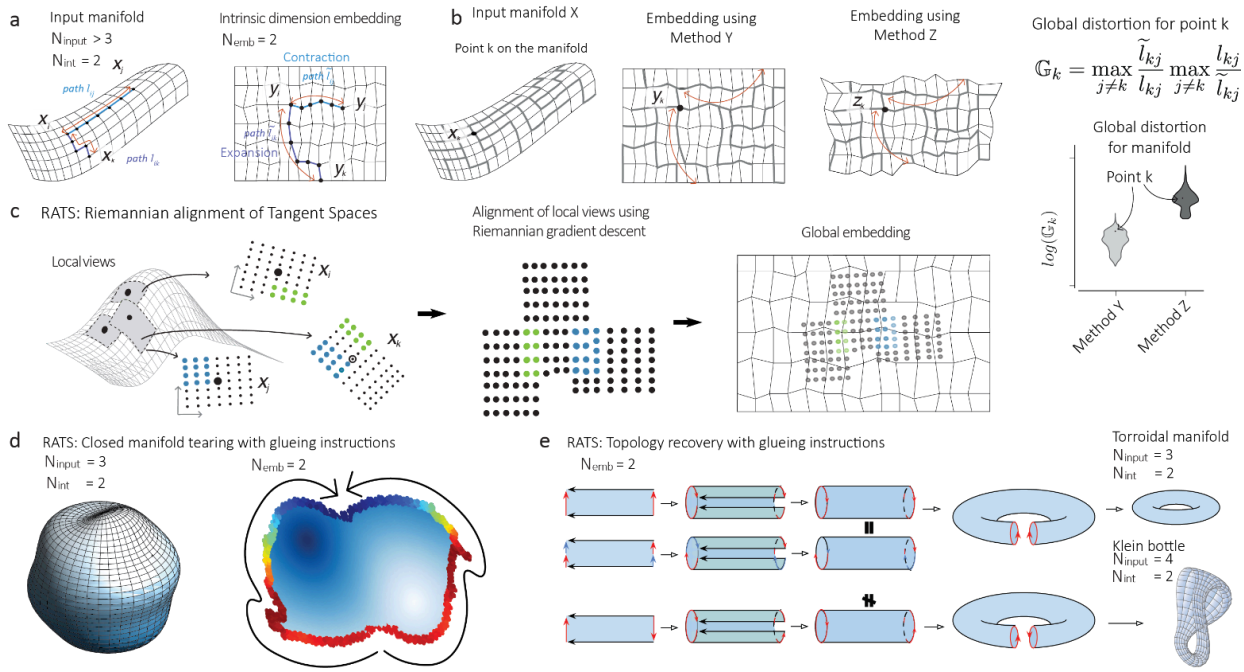

Figure 1: Distortion and RATS a) Embedding data points from an ambient dimension (N_{input}) into a low-dimensional embedding space (N_{emb}), which corresponds to the intrinsic dimension of the data (N_{int}). b) Embeddings can produce distortion due to contraction and expansion of relative distances along geodesic paths. Here, x_k represents data point k on the original manifold, and y_k and z_k represent its counterparts in two different embeddings. For each datapoint, we calculate its geodesic distances to all other points in the original and the embedding space, and identify the paths with the worst contraction (path is shorter in the embedding) and expansion (path is longer in the embedding) compared to the original space. The ratio of these defines the global distortion at a point. Calculating the distribution of global distortion for all points in the datasets, allows us to quantitatively evaluate which embedding better preserves path lengths from the original space. c) RATS computes low-distortion global embeddings in two phases. First low-dimensional low-distortion local views of the dataset are constructed. Second these are aligned, based on their overlaps (blue and green points), using Riemannian gradient descent to obtain a global embedding. d) RATS is capable of tearing closed/non-orientable manifolds in order to embed them into their intrinsic dimensions. The input is a closed surface in a three-dimensional space i.e. N_{input} = 3 and N_{int} = 2. RATS tears the input manifold to embed it into its intrinsic dimension and provides gluing instructions at the output by coloring the points along the tear so that the same colored points are to be stitched together. The arrows constituting the gluing diagram are manually drawn by tracking the variation of the colors. e) The schematic shows the recovery of the topology of a torus and a Klein bottle through different gluing diagrams on the same 2D embedding.

Manifold learning through Riemannian Alignment of Tangent Spaces (RATS):

We also present a new bottom-up manifold learning technique called *Riemannian Alignment of Tangent Spaces* (RATS, Figure 1c-e) that generates embeddings with low distortion. In this bottom-up approach, RATS first generates low-distortion local views of the data, whose construction can accommodate different similarity measures (*e.g.*, Euclidean, cosine) to obtain a local parametrization of each point. RATS then learns a rigid alignment (rotations and translations) between these views, based on overlapping points between pairs of views, using Riemannian Gradient Descent[28]. This yields a single global embedding of the data. A key insight here is that for bottom-up manifold learning, ensuring low local distortion in generating local views and aligning them optimally, *i.e.* with small error, will result in minimal accrual of distortion in the global embedding. Thus, while RATS does not optimize a cost function based on Global distortion $\mathbb{G}_k$ to obtain an embedding, the design of its construction process will result in a global embedding with low global distortion.

Furthermore, RATS provides two features that are absent from previously proposed manifold learning methods. RATS uses a unique algorithm that enables the tearing of closed manifolds in order to embed these in their intrinsic dimension[29] and it provides visual gluing instructions to piece the torn manifold into its original closed topology (Figure 1d). This can provide a valuable means to decode external variables directly from the intrinsic dimension of the data manifold, which is often more relevant than the low dimension which it may be represented in by a manifold learning method. For example, the intrinsic dimension of a circle is one, being parametrized by a single variable, the radial angle. Most manifold learning techniques, however, require two parameters to recover an underlying circle from data (*i.e.*, the sine and cosine of the angle). Similarly, one's location on Earth can largely be described by two coordinates, latitude and longitude, and an atlas or map of the Earth is a representation that tears the manifold surface to embed it in two dimensions. For many manifold learning techniques, the attempt to embed a sphere into two dimensions will result in a collapsed manifold, where for instance, the north and the south hemispheres lie on top of each other in 2D (Supplementary figure 1). In such a case, we obtain an undesirable many-to-one mapping (*i.e.*, both the north pole and south pole are mapped to the same point), where information that enables the recovery of the original points in data space is lost, thereby hindering our ability to interpret and decode such relationships. This also leads to distance distortion where points that are far from each other in the data space (*e.g.*, Japan and Australia in Supplementary figure 1) are represented as being close or similar to one another in the embedding. By

tearing manifolds and embedding them in their intrinsic dimension, RATS provides a more intuitive and straightforward approach to the interpretation of closed manifolds.

Secondly, RATS provides information that indicates the points along which the manifold has been torn and how to stitch it back together, so that its topology can be recovered (Figure 1e). For instance, a low-dimensional rectangular embedding may represent more than one plausible topological object, for example, a hyperplane, a torus, or a Klein bottle. RATS can disambiguate these cases by providing information that can lead to the recovery of these two later distinct topologies (Figure 1e), or in the case of the former indicate that the manifold has not been torn.

Results

The true intrinsic dimension of manifold objects is often challenging to ascertain in many empirical datasets, which are often plagued by inconsistencies, noise, and discontinuities, among other irregularities. An ideal low-distortion manifold learning technique would be able to perform well in a wide variety of circumstances, *i.e.*, when the ground-truth topology and intrinsic dimensions are unknown, or where the presence of perturbations shroud our understanding of the true topology of low-dimensional objects. We, therefore, first use the *Global distortion* measure $\mathbb{G}_k$ to compare RATS to a variety of other manifold learning techniques for idealized datasets where the generative properties of manifolds are well understood but may not necessarily reflect properties of real-world data. Second, we apply RATS and other manifold methods to high-dimensional closed topological datasets recently published in neuroscience to evaluate whether RATS can provide a more general approach to data decoding compared to the specific approaches used in these examples. Finally, we compare the *Global distortion* in embeddings from various methods where the object topology may be camouflaged behind nuisance features like noise, sparsity and inhomogeneous density. We undertake an amalgamated approach where we generate data manifolds using deep learning methods that mimic realistic data features, while also providing a decipherable topological structure that can be linked to phenotypes, behaviors, or other external variables and evaluate decoding performance across methods. Overall, we hypothesize that using a quantifiable measure of distortion and methods that achieve low distortion would not only furnish higher-quality embeddings for visualizations but may additionally reveal the principles by which such topologies represent variables in the external world.

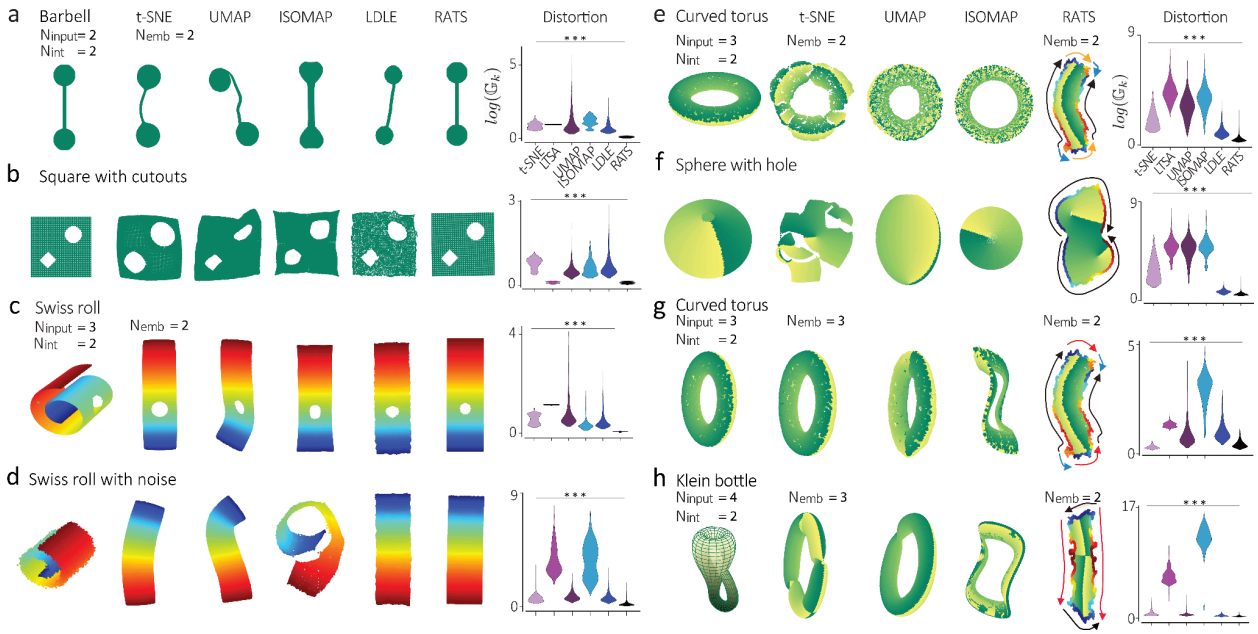

Figure 2: Distortion in embeddings of idealized datasets. a) Left: Barbell with $N_{input} \leq 2$ embedded in $N_{emb} = 2$ using t-SNE, UMAP, ISOMAP, LDLE, and RATS, where $N_{int} = 2$ (intrinsic dimension). Right: log G_k , Global distortion distributions for each embedding provided for t-SNE, LTSA, UMAP, ISOMAP, LDLE, and RATS (Stats N = 4929 Kruskal-Wallis test in Supplementary table 2). b) Square with cutouts with $N_{input} = 2$ embedded using the same techniques as a). Right: Same as a) for Square with cutouts. c) Swiss roll with circular cutout with $N_{input} = 3$ ($N_{int} = 2$) embedded in $N_{emb} = 2$ and global distortion distributions for each embedding (right). (Stats N = 5047 Kruskal-Wallis test in Supplementary table 2) d) Swiss roll with noise with $N_{input} = 3$ embedded in $N_{emb} = 2$, same as c). e) Left: Curved torus with $N_{input} = 3$ ($N_{int} = 2$) embedded in $N_{emb} = 2$ using t-SNE, UMAP, ISOMAP and RATS. Right: Global distortion distributions for the same techniques as in a) (Stats N = 5000 Kruskal-Wallis test in Supplementary table 2). f) Left: Sphere with a hole, embedded same as e) Right: same as e for Sphere with hole. g) Left: Curved torus with $N_{input} = 3$ ($N_{int} = 2$) embedded in $N_{emb} = 3$ using t-SNE, UMAP, ISOMAP and $N_{emb} = 2$ using RATS. Right: Global distortion distributions for embeddings in g) (Stats N = 5000 Kruskal-Wallis test in Supplementary table 2). h) Left: Klein bottle with $N_{input} = 4$ ($N_{int} = 2$), here shown in 3D, embedded using the same techniques as g). Right: same as e) for Klein bottle (Stats N = 4900 Kruskal-Wallis test in Supplementary table 2).

Evaluating RATS on idealized datasets

To evaluate the benefit of a low-distortion method like RATS in uncovering latent structure, we first consider idealized datasets where ground truth topologies are well-understood. We compute 2D (intrinsic dimension, $N_{\text{int}} = 2$) embeddings of manifolds with varying dimensionality (barbell, $N_{\text{input}} \leq 2$, Figure 2a), manifolds with missing data with distinct local geometries (square with cutouts, $N_{\text{input}} = 2$, Figure 2b), nonlinear manifolds with missing data (Swiss roll with a cutout, $N_{\text{input}} = 3$, Figure 2c) and with

noise (Swiss roll with noise, N_{input} = 3, Figure 2d). For each dataset, we calculate the distortion in embeddings produced by various techniques, *viz.*, t-SNE, UMAP, LTSA, ISOMAP, LDLE, and RATS. After optimizing hyperparameters for each technique (Table 1, Supplementary figure 2), a nonparametric analysis of variance (Kruskal-Wallis) reveals a significant main effect of global distortion across techniques for each of these datasets (Figure 2a-d, Supplementary table 1). RATS consistently exhibits the lowest global distortion[1] in each of these idealized datasets . We demonstrate that for these simple datasets, RATS-generated embeddings are most robust to noise and excel in preserving local geometric features, *i.e.*, the shapes of the cutouts, compared to other techniques.

Closed manifolds, such as a sphere or a torus, have no boundaries, and despite their intrinsic dimension being two, they are often processed in three or higher dimensions. RATS provides us with the ability to tear these manifolds and embed them into their 2D intrinsic dimension. Tearing closed manifolds may, however, exacerbate distortion as postulated by many classical results in differential geometry such as *Theorema Egregium*[30]. To study this, we first examine an idealized torus dataset and a sphere with a hole, and compare the optimized embeddings from various methods. We find that when we embed in intrinsic dimensions (N_{int} = 2) using conventional methods, many of the methods yield collapsed embeddings for both datasets, consequently incurring high distance distortion (Figure 2e,f). In contrast, RATS is able to provide a torn embedding whose gluing diagram indicates a toroidal and spherical topology, respectively (Figure 2e,f). Since other methods are unable to tear manifolds and provide collapsed embeddings, the question arises whether it is fair to compare these methods that cannot embed closed manifolds into intrinsic dimensions to RATS using distortion measures. To provide a more balanced comparison, we also compare results from embeddings of a curved torus by these methods in a non-intrinsic embedding dimension (N_{emb} = 3) and compare these to the RATS result in N_{emb}=N_{int} = 2 (Figure 2g). Even under these circumstances, RATS provides comparable or improved global distortion (Full statistical outcomes in Supplementary table 1).

Computer vision applications, such as network topologies of images[1] and conformation space of chemical molecules[31], have uncovered non-orientable manifolds like Klein bottles. This raises the question of how manifold learning techniques respond to non-orientable data manifolds. To address this, we examined whether various techniques could embed a four-dimensional Klein bottle dataset into its

[1] Note that RATS is not optimised on global distortion, which is an independent measure to evaluate the different manifold learning methods.

intrinsic dimension, N_{int} = 2 (Figure 2h). Most manifold learning techniques are better equipped to embed these data into three dimensions, which results in a pinch or a self-intersection, leading to high distortion. RATS, however, is able to embed the Klein bottle into its intrinsic dimension through tearing and provide a gluing diagram that correctly recovers its topology (Figure 1e, Figure 2f). Among these methods, t-SNE, UMAP, LDLE, and RATS provided relatively low-distortion embeddings with RATS providing the one with minimum distortion (Statistics in Supplementary table 1).

Evaluating RATS on biological datasets

Next, we examine whether biological data resembling closed manifolds can be described using the same manifold learning principles. We turn to two recent studies in neuroscience that unveil intriguing topological manifolds in the neural population representations for head-direction[12,32] and navigation[13] systems in mammals. In the first study, the authors examined neural activity from head direction cells in the anterodorsal thalamic nucleus (ADn) of a mouse while it forages in an open 2D environment. They then embed the 22-dimensional head direction activity dataset[12,32], which exhibits a closed ring-like structure (Figure 3a). Although the intrinsic dimension of the ring is one, several conventional manifold learning techniques produce a representation of this topology[24] that requires at least two dimensions to produce a low-distortion parameterization of the data. To address this, the authors[12,30] developed a specialized approach based on spline fitting to obtain a one-dimensional parameterization of the neural activity as it corresponds to heading direction. Here, we embed the neural activity in N_{emb} = 2 dimensions using t-SNE, LTSA, UMAP, and ISOMAP and in N_{emb} =N_{int} = 1 dimension using RATS (Figure 3b). The gluing instructions produced by the RATS embedding depicts a closed-loop topology. This approach may prove more generally applicable compared to the spline fitting approach since it does not mandate a one-dimensional closed manifold. We decode the head direction angle from LDLE and RATS embeddings by simply scaling the one-dimensional embeddings to range between 0 and 2π, while for other methods, we employ the spline fitting approach[12] on the two-dimensional embeddings (Figure 3c-d, Supplementary figure 3a-f). We assess the mean squared error between the decoded head direction angle from the embeddings and the angle measured by a light-emitting diode tracking (Figure 3c-d). The RATS embeddings result in the least error and these results generally hold with hyperparametric variation (Supplementary Figure 4).

In the second study, the authors investigated neural activity from grid cells in the medial entorhinal cortex (MECII) of rats, revealing that the activity patterns collectively reside on a toroidal manifold[13], with

the rat's movements in the environment corresponding to trajectories on the torus (Figure 3e). We compare manifold learning techniques on the 149-dimensional neuronal recordings reduced to six dimensions via PCA (see methods for preprocessing of the data) to recapitulate the reported findings using UMAP with $N_{emb} = 3$ (Figure 3f). Since the intrinsic dimension of the toroidal object is two, we employ RATS to embed the data into its intrinsic dimensions. The resulting gluing diagrams depict a toroidal topology (Figure 3f, Supplementary figure 5). We further went on to quantify the topology using persistence barcodes computed from the shortest path distances on the data and the embeddings. It is worth noting that only the tear information, *i.e.* the points along the tear, from the RATS embedding was utilized to compute the 1-homology, leading to a faster computation. The resulting persistence barcodes align with the ones derived directly from the data (Figure 3g, Supplementary figure 8). Additionally, we compared the spatial firing rate maps with the firing rate maps on the RATS embedding (Figure 3h, Supplementary figure 7). We emphasize that RATS provides the means to directly embed in the intrinsic two dimensions, therefore precluding the need for extensive computation of cocycles, as employed in previous work[13]. Depending on the application, the RATS embedding could thus complement Topological Data Analysis (TDA) algorithms that determine the homology of the data manifold to infer its topology.

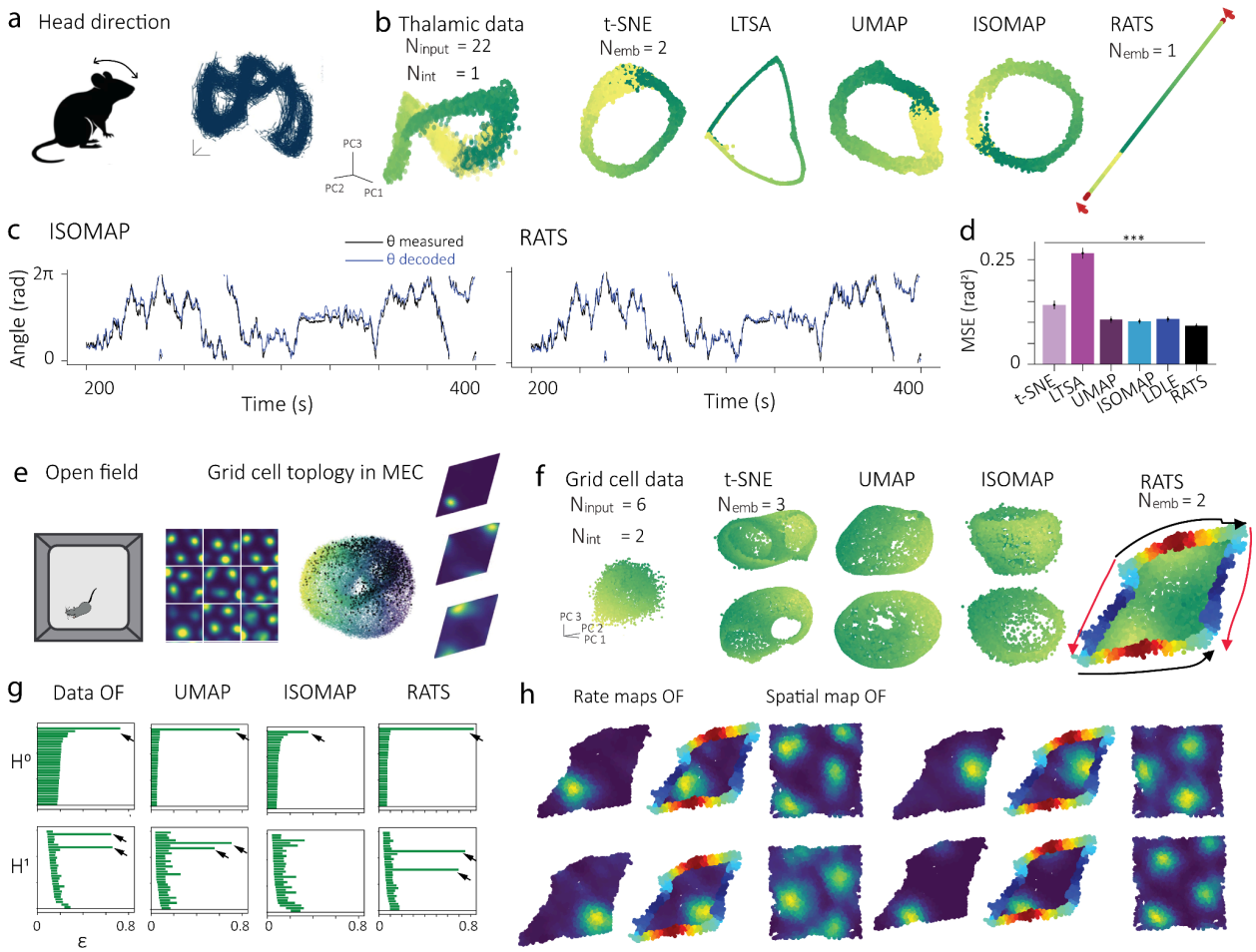

Figure 3: *Neural activity from head direction cells and grid cells. a) The 22-dimensional population activity vectors recorded from the anterodorsal thalamic nucleus (ADn) of mice while foraging in an open 2D environment have been shown to organize on a 1D manifold, visualized in 3D with ISOMAP[12]. b) PCA visualization of the input neural activity in 3D. Embeddings of the neural activity are obtained using t-SNE, LTSA, UMAP, and ISOMAP ($N_{emb} = 2$) and using RATS ($N_{emb} = N_{int} = 1$). RATS provides gluing instructions by coloring the ends of the 1D embedding, indicating a ring structure. c) Unsupervised decoding of the head direction angle from the RATS embeddings by scaling the one-dimensional embedding to range between 0 and 2π. For all other methods, decoding is performed using a spline parameterization of the embeddings as in ref [12]. The decoded angle is compared against the direct measurement of the head orientation of the animal from light-emitting diode tracking. d) RATS exhibits the lowest mean squared error (MSE) between the decoded angle due to each method and the measured angle. e) The 149-dimensional population activity vectors recorded from the medial entorhinal cortex (MECII) of rats while foraging in an open 2D environment have been shown to organize on a toroidal manifold, visualized in 3D with UMAP[13]. f) Visualizing the input neural activity in 3D using PCA. Embeddings of the 6-dimensional PCA reduced neural activity are obtained using t-SNE, LTSA, UMAP, and ISOMAP ($N_{emb} = 3$) and using RATS ($N_{emb} = N_{int} = 2$). The tear on the RATS embedding is regularized using a graphical user interface (GUI) (see Methods). RATS provides*

gluing instructions by coloring the two sides of the 2D embedding. The derived gluing diagram depicts toroidal topology (Figure 1e). g) The persistence barcodes for zeroth H_0 and first H_1 cohomology of the input data and the embeddings (see Methods). h) Visualization of the firing rate maps of four of the grid cells on the RATS embedding and in the spatial coordinates.

Evaluating RATS on manifolds generated by deep networks:

To understand the impact of distortion on the decoding of features in artificial and biological settings, we examine datasets whose generative properties are well-characterized, however, they exhibit features that resemble real-world datasets. For this purpose, we use deep recurrent networks to generate rich and organized topologies that resemble manifolds derived from the dynamics of previously-reported activity of neural populations[33–35]. We use recurrent neural networks (RNNs) to generate latent dynamics that yields a variety of topological objects, where we understand how latent variables along these manifolds relate to the networks' behavioral outputs. Furthermore, we study scenarios where the manifold geodesics exhibit properties that can be expected from realistic datasets, such as, variable density, missing or noisy data, discontinuities and tearing of the manifold structure.

We first examine the influence of varying the density across the data points on manifold learning methods. We then quantify the distortion induced by such variation and its impact on decoding behavioral metrics. In a motor timing study[13,33], animals and RNNs were trained to use contextual information (Context) to generate different time intervals (Response) from the onset of a Start cue (Figure 4a,c). By recreating network activity from this task, we were able to generate a three-dimensional manifold (with N_{int} = 2) representing the evolution of network dynamics from the Start to Response duration (Figure 4b left). The manifold was organized to reflect the logic of contextual inputs, where the speed of the trajectory along the manifold was inversely related to the time interval generated by the network, recapitulating the original finding using data from the macaque medial frontal cortex (Figure 4c, Supplementary figure 9a). The ensuing manifold contained varying densities of points along both intrinsic dimensions. We compare the embeddings of various techniques and calculate global distortion for each case (Supplementary figure 10). In a non-parametric analysis of variance, we find a significant main effect of global distortion, with RATS providing the lowest distortion embeddings for these datasets (Figure 4b right, Statistics in Supplementary table 2 and 3; Supplementary figure 9b compares ISOMAP and RATS). In addition, we find that when decoding environmental variables from metrics derived from the embeddings such as the response time from the neural trajectory speed, the quality of decoding

corresponded well with the low-distortion nature of the embedding (Supplementary figure 9a). We found that RATS and ISOMAP, which provide the lowest distortion embeddings, also outperformed other manifold learning techniques in decoding for this dataset, of which RATS provides the lower distortion (Supplementary figure 9b).

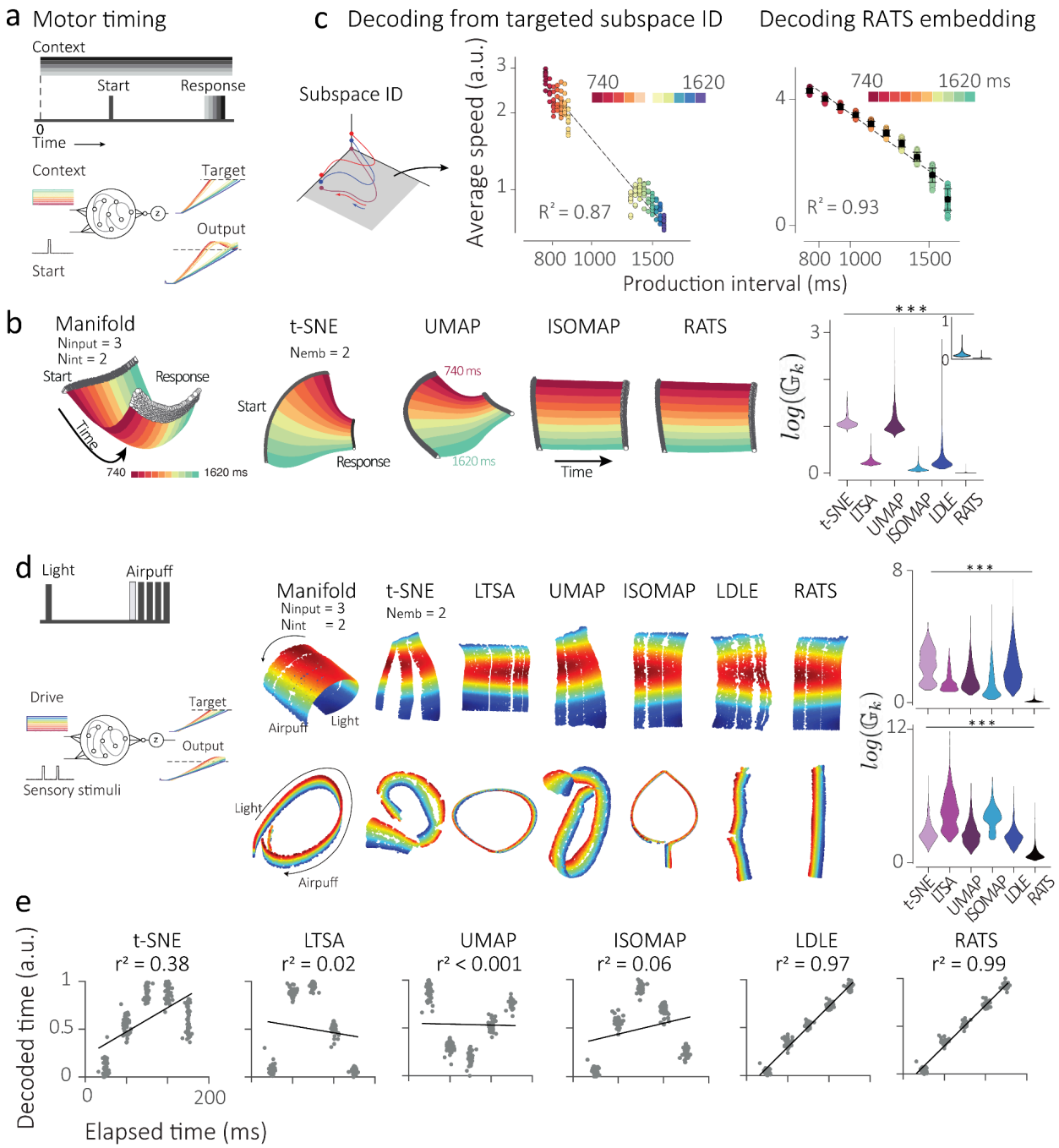

Figure 4. Embedding irregular manifolds generated through deep networks. a) Based on ref[33], RNNs were trained to associate context cues with intervals by ramping to a threshold at different times. b) Left: Nonlinear manifold obtained from the network dynamics (left) embedded using t-SNE, UMAP, ISOMAP and RATS. Colors represent context-dependent time interval associations (red to green: 740 - 1620 ms). Right: Global distortion distributions for embeddings of the nonlinear motor timing manifold using t-SNE, LTSA, UMAP, ISOMAP, LDLE and RATS. Non-parametric analysis of variance using Kruskal-Wallis . c) Left: In ref [33], neural population activity speeds for different contexts correlated with the produced time interval (red to violet: 740 - 1620 ms). Right: Nonlinear manifolds generated by dynamics of trained RNNs recapitulate this principle, where despite variation in density of the manifold for short and long durations, decoding of the manifold yields a high correlation with produced time intervals. d) Left: Manifolds generated by RNNs trained on an associative learning task between two sensory signals (light and airpuff)[36]. RNNs are provided a continuous driving signal that represents fluctuations in initial condition of the dynamics. Resulting nonlinear manifolds (N_{input} = 3) containing missing data and containing a forked tail are embedded using t-SNE, LTSA, UMAP, ISOMAP, LDLE and RATS. Right: Global distortion distributions for each embedding for the manifolds. e) Decoding of time intervals from the embeddings by various techniques reveals a linear relationship with elapsed time for LDLE and RATS.

Noise is a frequently encountered nuisance variable in biological datasets[12,13,32,33]. To examine the robustness of manifold learning techniques to off-manifold and on-manifold perturbations, Gaussian noise was added to the input or to each network unit, resulting in exacerbated density variations within and in the vicinity of the original data manifold. We again found that ISOMAP and RATS outperform the rest when dealing with such data, with RATS providing a lower distortion of the two (Supplementary figure 9c).

In a second example, we examine features of realistic datasets that often induce distortion into embeddings generated by many manifold learning techniques, such as variation in the density of data points, discontinuities in the data manifold, and fraying of manifold structure. To generate these conditions, we use a Pavlovian associative learning paradigm (Figure 4d), which is widely studied in neuroscience and behavioral science. Here, animals learn anticipatory behaviors by learning associations between two sensory cues. Neural activity is reported to briefly modulate during the delay period before returning to its baseline state[36], suggesting that the population activity is likely to manifest along a circular trajectory. Owing to fluctuations and noise, however, it may not always return to the same baseline value[35]. As a result, the population activity during the task resembles irregular Swiss roll-like objects, which vary in density, may be discontinuous, and may fray at locations. We embed two

network-generated variants of this topology during this behavior using various techniques (Figure 4e right). We find that while most techniques were able to embed these data manifolds into two dimensions (Figure 4d, Supplementary figure 11), they introduce significant distortions through tearing (t-SNE), non-uniform expansion (LTSA, UMAP), and overall warping (ISOMAP, LDLE). RATS is, however, able to preserve the discontinuities in the data manifold and provide the lowest distortion embedding among these (Figure 4d, Statistics in Supplementary table 3). In another variant, we use a data manifold that exhibits fraying and sparse sampling along the manifold (Figure 4e below). Under these circumstances, many techniques are unable to successfully embed the unraveled structure in 2D (t-SNE, LTSA, UMAP, ISOMAP) and exhibit distortions at the frayed end (t-SNE, UMAP, ISOMAP). Again, RATS provides the lowest-distortion embedding among these (Figure 4e).

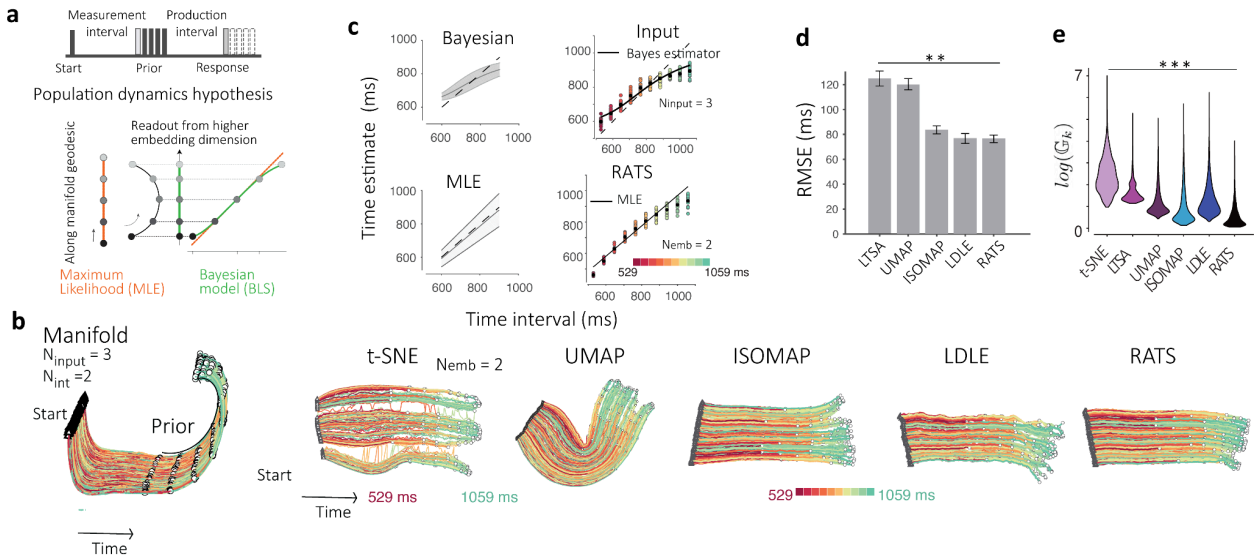

Figure 5: Hypothesis validation using manifold learning techniques. a) For a Bayesian interval reproduction task, temporal stimuli are sampled from a uniform prior distribution, which causes neural and network population trajectories to curve in a semicircular trajectory. Linear readouts of these populations lead to states that become compressed towards the mean of the prior, leading to biased estimates towards the mean of the prior, a hallmark of Bayesian inference using prior knowledge (BLS, green). Removal of the influence of the prior (orange), should lead to a mitigation of the bias but an increase in variability and the maximum likelihood estimator (MLE). b) Left: Nonlinear manifold formed by network dynamics in the temporal reproduction task exhibits time intervals along a semicircular curvature of the population manifold (N_{input} = 3). Right: Embeddings into N_{int} = 2 using t-SNE, UMAP, ISOMAP, LDLE and RATS. c) Left: Normative model of Bayes Least Squares estimation of time intervals and Maximum likelihood estimation. Right: Linear decoding yields estimates consistent with BLS. Linear decoding of the

embedded manifold by RATS yields estimates consistent with MLE. d) Comparison of RMSE of decoded estimates from the stimuli ensuing from embeddings with different techniques. Error bars represent standard deviation. e) Global distortion distributions for the optimized embeddings provided by each technique.

In addition to decoding metrics and providing data visualization, accurate low-distortion embeddings could enable the validation or invalidation of hypotheses in computational biology and other fields. Here we provide an example from neuroscience where recent proposals[29,37] suggest that manifold-like topologies in neural population structure may enable different computations based on how information is decoded from the manifold, *i.e.*, linear projections of the manifold embedded in high-dimensional space may serve some computations whereas, other computations may be more suited to readouts like geodesic distances along the manifold's intrinsic dimensions. For example, recent work[34] on time interval reproduction demonstrates that neural activity curves in semicircular trajectories (Figure 5a), potentially represent prior knowledge of time distributions that bias behavioral estimates in a manner consistent with Bayesian theory (Figure 5a). If prior knowledge is hypothesized to be encoded as curvature in such geometries, then unfolding such structures and decoding them along their intrinsic dimension should yield temporal estimates that are free of the influence of prior knowledge[37], *i.e.*, consistent with non-Bayesian models like Maximum likelihood estimates (Figure 5b,c). We find that manifolds representing such prior-related curvature, whose linear readouts are hypothesized to yield Bayes-like estimates (Figure 5c left), when embedded in lower dimensions and decoded, do indeed yield values that are consistent with Maximum likelihood estimation (Figure 5c , Supplementary figure 11a-c). Of the various techniques we tested, RATS provides the lowest distortion embeddings (Figure 5e). The quality of manifold decoding can often depend on metrics that are selected along intrinsic dimensions with different methods offering different trade-offs. We show that RATS consistently provides one of the lowest overall root mean squared error (RMSE) for different measures of decoding of time intervals from latent intrinsic variables (Supplementary figure 11a-c) and global distortion in its embedding (Supplementary figure 12). Therefore, low-distortion embeddings can be used for hypothesis validation and for reliable decoding of relationships between latent low-dimensional variables and observable external variables.

Discussion

With the emergence of low-dimensional topological and geometric structures in complex multivariate datasets across biological and non-biological fields, it becomes increasingly crucial to accurately isolate

these structures into their native intrinsic state with minimal distortion to draw accurate scientific conclusions. To address this need, RATS serves as an unsupervised manifold learning technique designed to embed latent topological structures, including closed and non-orientable manifolds, into intrinsic dimensions with low distortion. We demonstrate its consistency in generating low-distortion embeddings across a wide variety of datasets, including data with complex topologies and noisy biological datasets with unknown ground truth. Overall, we consider RATS a generalist technique that especially mitigates the ubiquitous distortion problem encountered in high-dimensional data science and performs well under diverse real-world data scenarios.

We have shown that RATS effectively embeds data in its intrinsic dimension, but an important question is how to determine that intrinsic dimensionality. Many times, the experiment has a controlled and known number of variables but in other settings, for exploratory data analysis, manifold embeddings can reveal additional unconsidered variables or metadata that could be represented. In such cases, the intrinsic dimensionality can be estimated from the data itself, for example estimating the fractal dimension via the correlation dimension[38,39], or doubling dimension via the Levina–Bickel Maximum likelihood estimator [40]. For a model-driven approach, one can run the embedding method for an increasing number of dimensions and determine whether increasing the embedding dimensionality leads to a significant decrease in optimization loss, e.g., the Kullback–Leibler (KL) divergence for t-SNE, or for spectral approaches by evaluating the spectrum of eigenvalues, *e.g.*, in PCA[41].

The evaluation of embedding techniques is an open and debated question, amongst others, most recently in bioinformatics for 2D visualization of, *e.g.*, single-cell RNA sequencing data. Multiple measures have been proposed to quantify local or global structure, where for some measure the underlying assumption of the data model is that it is formed of clusters[17,42–45]. These evaluation criteria mostly rely on supervised information, *i.e.*, knowledge of the labels of clusters in the data, for example to evaluate the performance of downstream analysis on the embedding. However, such evaluation is not possible in the absence of such ground-truth information or prior knowledge of the data. In machine learning, various unsupervised distortion measures, *e.g.*, average and worst-case local distortion, have been shown to vary in their ability to quantify ranking of neighbors and robustness to noise and outliers[27].

We introduced a novel measure of distortion, called global distortion, which relies on calculating the geodesics between points, which enables the preservation of *geometry* of the data in the embedding space. In general this is important for quantifying distortion when data are not embedded into Euclidean space, but into alternative geometries, *e.g,.* by constraining the representation to lie on a sphere[26] or in hyperbolic space, which is used for embedding hierarchical data. For noisy data, estimating geodesics can be susceptible to "shortcuts" between points, however, this can be mitigated by proper preprocessing. Our global distortion measure enables us to construct a distribution of distortion across the entire datasets for a given low-dimensional embedding, thereby allowing us to evaluate the quality of the embedding for different methods. None of the manifold learning methods presented here are optimized on this measure. We see that different methods can generate low-distortion embeddings for specific dataset features, but have poor quality for others. We do find, however, that although it does not always provide a uniquely low-distortion solution, RATS consistently provides low global distortion distributions for a wide variety of data contingencies in biological, synthetic, and idealized datasets.

Furthermore, we demonstrate the utility of low-distortion embeddings in linking external variables to latent structures on the manifold and for validating hypotheses on biological computations. As an important example of this, our experiments primarily focused on data whose parametrization of the underlying manifold was a physical quantity, *i.e.* space in the navigation circuit[13,12] or time in motor control[33]. Our findings highlight the implicit link between distortion in embeddings and data analysis based on these embeddings: reducing distance distortion in low-dimensional representations of data leads to improved downstream analysis, *e.g.* improved decoding of external variables. For this, we both demonstrated that embeddings with high distortion had the highest decoding errors, as well as showed that RATS achieved the lowest distortion while incurring the lowest error in most cases. By minimizing distortion we better reveal how real physical quantities are represented in latent low-dimensional representations of large datasets, since discrepancies in the representations of these quantities will correspond to real phenomena rather than artifacts introduced by the dimensionality reduction approach. Thus, quantifying distortion and reducing distortion in embeddings may be essential to distinguish whether given information is truly absent from the data or whether erroneous representation owing to distortion has masked our ability to access said information. It remains an interesting question how distortions affect latent representations of more abstract stimuli or complex behaviors of natural and artificial systems.

References:

1. Carlsson, G. & Gabrielsson, R. B. Topological Approaches to Deep Learning. *Topological Data Analysis* 119–146 (2020).
2. Turian, J., Ratinov, L. & Bengio, Y. Word representations: a simple and general method for semi-supervised learning. in *Proceedings of the 48th annual meeting of the association for computational linguistics* (2010).
3. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
4. Fefferman, C., Mitter, S. & Narayanan, H. Testing the manifold hypothesis. *Journal of American Mathematical Society* (2016).
5. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
6. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716–729.e27 (2018).
7. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
8. Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).
9. Nicolau, M., Levine, A. J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7265–7270 (2011).
10. Jackson, M. D., Xu, H., Duran-Nebreda, S., Stamm, P. & Bassel, G. W. Topological analysis of multicellular complexity in the plant hypocotyl. *Elife* **6**, (2017).
11. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural Manifolds for the Control of Movement. *Neuron* **94**, 978–984 (2017).
12. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat. Neurosci.* **22**, 1512–1520 (2019).
13. Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).
14. Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron* **98**, 1005–1019.e5 (2018).
15. Beiran, M., Meirhaeghe, N., Sohn, H., Jazayeri, M. & Ostojic, S. Parametric control of flexible timing through low-dimensional neural manifolds. *Neuron* **111**, 739–753.e8 (2023).
16. Brennan, C. & Proekt, A. A quantitative model of conserved macroscopic dynamics predicts future motor commands. *Elife* **8**, (2019).
17. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
18. Tamazian, G. *et al.* t-SNE Highlights Phylogenetic and Temporal Patterns of SARS-CoV-2 Spike and Nucleocapsid

Protein Evolution. in *Bioinformatics Research and Applications* 255–262 (Springer Nature Switzerland, 2022).

19. Wang, S.-C. *et al.* Novel Imaging Revealing Inner Dynamics for Cardiovascular Waveform Analysis via Unsupervised Manifold Learning. *Anesth. Analg.* **130**, 1244–1254 (2020).
20. Shen, C., Lin, Y. T. & Wu, H. T. Robust and scalable manifold learning via landmark diffusion for long-term medical signal processing. *J. Mach. Learn. Res.* (2022).
21. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).
22. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] (2018) [arXiv].
23. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
24. Zhang, Z. & Zha, H. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. *SIAM J. Sci. Comput.* **26**, 313–338 (2004).
25. Kohli, D., Cloninger, A. & Mishne, G. LDLE: Low Distortion Local Eigenmaps. *J. Mach. Learn. Res.* **22**, (2021).
26. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).
27. Chennuru Vankadara, L. & von Luxburg, U. Measures of distortion for machine learning. *Adv. Neural Inf. Process. Syst.* **31**, (2018).
28. Absil, P.-A., Mahony, R. & Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. (Princeton University Press, 2009).
29. Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol.* **70**, 113–120 (2021).
30. Oprea, J. *Differential Geometry and Its Applications*. (MAA, 2007).
31. Scoccola, L. & Perea, J. A. FibeRed: Fiberwise dimensionality reduction of topologically complex data with vector bundles. arXiv:2206.06513 [cs.CG] (2022) [arXiv].
32. Peyrache, A., Lacroix, M. M., Petersen, P. C. & Buzsáki, G. Internally organized mechanisms of the head direction sense. *Nat. Neurosci.* **18**, 569–575 (2015).
33. Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* **21**, 102–110 (2018).
34. Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian Computation through Cortical Latent Dynamics. *Neuron* **103**, 934–947.e5 (2019).
35. Narain, D., Remington, E. D., Zeeuw, C. I. D. & Jazayeri, M. A cerebellar mechanism for learning prior distributions of time intervals. *Nat. Commun.* **9**, 469 (2018).
36. De Zeeuw, C. I., Koppen, J., Bregman, G. G., Runge, M. & Narain, D. Heterogeneous encoding of temporal stimuli in the cerebellar cortex. *Nat. Commun.* **14**, 7581 (2023).
37. Sohn, H. & Narain, D. Neural implementations of Bayesian inference. *Curr. Opin. Neurobiol.* **70**, 121–129

(2021).

38. Grassberger, P. & Procaccia, I. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena. *Physica D* **9**, 180–208 (1983).
39. Kalantana, J. & Einbecka, Z. Intrinsic dimensionality estimation for high-dimensional data sets: New approaches for the computation of correlation dimension. *J. Emerg. Technol. Web Intell* **5**, 91– (2013).
40. Levina, E. & Bickel, P. Maximum likelihood estimation of intrinsic dimension. *Adv. Neural Inf. Process. Syst.* **17**, (2004).
41. Nguyen, L. H. & Holmes, S. Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* **15**, e1006907 (2019).
42. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLoS Comput. Biol.* **19**, e1011288 (2023).
43. Espadoto, M., Martins R M Kerren, T., H. N. S. & Telea, A. C. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics* **27**, 2153–2173 (2021).
44. Huang, H., Wang, Y., Rudin, C. & Browne, E. P. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications Biology* **5**, 719.
45. Wang, K. *et al.* Comparative analysis of dimension reduction methods for cytometry by time-offlight data. *Nat. Commun.* **14**, 1835 (2023).
46. Bhatia, N. & Vandana. Survey of Nearest Neighbor Techniques. *arXiv e-prints* arXiv:1007.0085 (2010).
47. Kohli, D., Mishne, G. & Cloninger, A. Non-degenerate Rigid Alignment in a Patch Framework. arXiv:2303.11620 [math.NA] (2023) [arXiv].
48. Kloke, J. & Carlsson, G. Topological De-Noising: Strengthening the Topological Signal. arXiv:0910.9547 [cs.CG] (2009) [arXiv].
49. Bauer, U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology* **5**, 391–423 (2021).
50. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
51. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
52. Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent Network Models of Sequence Generation and Memory. *Neuron* **90**, 128–142 (2016).
53. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
54. Rajan, K. & Abbott, L. F. Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.* **97**, 188104 (2006).
55. Werbos, P. J. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. (John Wiley & Sons, 1994).

56. Jazayeri, M. & Shadlen, M. N. A Neural Mechanism for Sensing and Reproducing a Time Interval. *Curr. Biol.* **25**, 2599–2609 (2015).
57. Jazayeri, M. & Shadlen, M. N. Temporal context calibrates interval timing. *Nat. Neurosci.* **13**, 1020–1026.(2010).

Acknowledgements

The authors thank Juan Alvaro Gallego and Mikio Aoi for their comments on the manuscript.

Funding:

This work was supported by funding from the Kavli Institute for Brain and Mind (DK), NIH (EB026936 to GM and DK), the NSF (CCF-2217058 to GM, DMS-2012266 to AC), a gift (Intel to AC), the Netherlands Organization for Scientific Research (Vidi-VI.193.076, Aspasia-015.016.012 and Gravitation-024.005.022 to DN), AiNed foundation (DN) and the NGF (NGF.1609.241.021 to DN).

Author Contributions

DK, JSN, GM and DN conceived the study. All authors developed the analytical approach. DK and JSN implemented the methods and carried out all analyses. All authors wrote the manuscript. GM and DN supervised this project and provided funding for it.

Competing Interests

The authors declare no conflicts of interest exist.

Data and materials availability

Our generated datasets will be available at the conclusion of peer-review.

Our code will be made available at an open access repository at the conclusion of peer-review.

Methods:

RATS Algorithm:

Given n data points, the algorithm consists of three steps: (i) computation of n low-distortion local parameterizations that map the local neighborhoods on the dataset into lower dimensions, yielding a local view of the data. (ii) merge local views into $m \ll n$ intermediate parameterizations with larger neighborhoods, yielding intermediate views of the dataset. (iii) These views are aligned to obtain a global embedding by minimizing a quadratic over the product of orthogonal groups using Riemannian gradient descent. These steps are described in detail below.

Comparison to other bottom-up manifold learning techniques

RATS belongs to a family of bottom-up manifold learning techniques, which first construct local parameterization of the data, i.e. every point has its own mapping so that its local neighborhood has a low-dimensional and a low-distortion representation. Both LTSA and RATS use a form of local PCA for this step, while LDLE uses a spectral construction procedure that selects (for each neighborhood) subsets of global eigenvectors of the graph Laplacian representation of the data. In the subsequent global alignment step, these methods aim to find transformations of each local parametrization, so that the representations of the overlap between neighborhoods, due to the respective (transformed) parameterizations, are as close to one another as possible. This corresponds to minimizing the alignment error between transformed local views. The global alignment step in LTSA assumes a uniform distribution on features vectors such that the global transforms can be affine, thus limiting its capacity to preserve the aspect ratio of the manifold, and often leading to highly-distorted embeddings. On the other hand, RATS makes no such assumptions about the data, using rigid transforms instead, which preserves distances between points, and is key to its achieving low-distortion embeddings. LDLE similarly uses rigid transforms, however it uses a heuristic Procrustes-based method to calculate the global embedding, and this in addition to its local view construction, leads to higher distortion than RATS.

Local Parameterizations:

Given a dataset $(x_k)_1^n \subseteq \mathbb{R}^p$ and an embedding dimension $d \leq p$, we start by computing a local neighborhood $\mathcal{N}_k$ around each point (using a nearest neighbor algorithm[46]). A local parameterization $\Phi_k : \mathbb{R}^p \to \mathbb{R}^d$, for the neighborhood $\mathcal{N}_k$ of x_k is then obtained by applying kernel PCA on the neighborhood *i.e.*, on $\{x : x \in \mathcal{N}_k\}$, with an appropriate linear or nonlinear kernel $K(\cdot, \cdot)$. In particular,

if $K(x,y) = \phi(x)^T \phi(y)$ for a certain choice of ϕ, then the kth local parameterization is given by $\Phi_k(x) = P_k \phi(x) + \mu_k$ where the d rows of P_k contain the d leading eigenvectors of the covariance of $\{\phi(x) : x \in \mathcal{N}_k\}$, and μ_k represents the centroid of the same collection. In our experiments, we utilized the linear kernel $K(x,y) = x^T y$ so that $\phi(x) = x$ and the cosine kernel $K(x,y) = x^T y / \|x\|_2 \|y\|_2$ where $\phi(x) = x / \|x\|_2$. In practice, the explicit value of $\phi(x)$ is not needed since the calculations involve dot products that can be directly evaluated using the kernel function.

<u>Local Distortion:</u>

Local distortion is calculated as the product of maximum expansion and maximum contraction of pairwise distances from the high dimensional data to the low dimensional embedding, where we use Euclidean distances in both spaces. Let Ψ denote a generic local parameterization on a set of points $\mathcal{U}$, local distortion is defined as,

$$\text{Local-Distortion}(\Psi, \mathcal{U}) = \max_{\substack{x_j, x_{j'} \in \mathcal{U} \\ x_j \neq x_{j'}}} \frac{\|\Psi(x_j) - \Psi(x_{j'})\|_2}{\|x_j - x_{j'}\|_2} \max_{\substack{x_j, x_{j'} \in \mathcal{U} \\ x_j \neq x_{j'}}} \frac{\|x_j - x_{j'}\|_2}{\|\Psi(x_j) - \Psi(x_{j'})\|_2}.$$

Then the distortion incurred by the kth local parameterization on the k' the neighborhood is obtained by substituting Φ_k for Ψ and $\mathcal{N}_{k'}$ for $\mathcal{U}$ *i.e.* $\text{Local-Distortion}(\Phi_k, \mathcal{N}_{k'})$.

<u>Postprocessing Local Parameterizations:</u>

To remove the anomalous local parameterizations which incur abnormally high local distortion, a postprocessing step[25] is performed by repeatedly replacing a parameterization Φ_k of the neighborhood $\mathcal{N}_k$ of x_k with that of a neighbor $\Phi_{k'}$ where $x_{k'} \in \mathcal{N}_k$, that produces the least local distortion on $\mathcal{N}_k$ *i.e.*, $k' = \text{argmin}_{j: x_j \in \mathcal{N}_k} \text{Local-Distortion}(\Phi_j, \mathcal{N}_k)$. This procedure is repeated until no further replacement is possible.

Intermediate Parameterizations and Views:

Each of the n local parameterizations Φ_k yields a local view $\Phi_k(\mathcal{N}_k) = \{\Phi_k(x) : x \in \mathcal{N}_k\}$. A large number of small local views lead to high computational cost and instability during their global alignment[25]. Such instability is due to ambiguous alignments caused by small overlap between the small neighborhoods[47]. Therefore, through a clustering procedure we obtain $m \ll n$ intermediate neighborhoods $\mathcal{V}_i$ yielding intermediate views $\widetilde{\Phi}_i(\mathcal{V}_i) = \{\widetilde{\Phi}_i(x) : x \in \mathcal{V}_i\}$, as follows. Beginning with each point belonging to its own cluster with an associated parameterization, the clustering procedure moves points into clusters based on the following rules: (i) a point can only move into one of its neighbors' clusters, (ii) it cannot move into a cluster that is smaller than the cluster it currently belongs to, and (iii) among the remaining potential clusters it will move to the one whose parameterization results in the least local distortion. This procedure is a greedy iterative procedure to merge the points into clusters. After pruning the empty clusters, the parameterizations associated with the remaining clusters are denoted $\widetilde{\Phi}_i$. The points in the ith cluster are given by $\mathcal{C}_i$ and the cluster label of the kth point is given by c_k. Finally, the intermediate neighborhood $\mathcal{V}_i$ associated with the *i*th intermediate view $\widetilde{\Phi}_i$ is the union of the neighbors of the points in of cluster $\mathcal{C}_i$ *i.e.*, $\cup_{x_k \in \mathcal{C}_i} \mathcal{N}_k$.

Aligning Intermediate Views for a Global Embedding using Riemannian Gradient Descent:

To obtain a low-distortion global embedding, we find rigid alignments between the intermediate views. For convenience, we define the parametrization of the *k*th point in the *i*th intermediate view as $x_{k,i} = \widetilde{\Phi}_i(x_k)$ and, Γ to be a bipartite graph between n points and m views. An edge in Γ is given by a tuple i.e. $(k,i) \in E(\Gamma)$ if $x_k \in \mathcal{V}_i$. Then the aim of rigid alignment is to find m orthogonal matrices $(S_i)_1^m$ and translation vectors $(t_i)_1^m$, that minimizes the alignment error,

$$\sum_{(k,i),(k,j) \in E(\Gamma)} \left\| (S_i^T x_{k,i} + t_i) - (S_j^T x_{k,j} + t_j) \right\|_2^2$$

so that on average after alignment, the parametrizations of a point in two views containing the point are as close as possible.

An equivalent problem is

$$\min \left\{ \sum_{(k,i) \in E(\Gamma)} \left\| y_k - (S_i^T x_{k,i} + t_i) \right\|_2^2 : (S_i)_1^m \subseteq \mathbb{O}(d), (t_i)_1^m, (y_k)_1^n \subseteq \mathbb{R}^d \right\}. \quad (1)$$

Following further reformulation of this problem as a quadratic program in the orthogonal matrices $(S_i)_1^m$, we obtain an approximate solution using Riemannian gradient descent which is guaranteed to converge locally linearly to a non-degenerate optimal alignment[47].

Note that LTSA uses *affine* transformations for global alignment and assumes that the feature vectors are uniformly distributed in the embedding *i.e.,* $YY^T = I_d$ where $Y = [y_1, \ldots, y_n]$, and y_k is the final global embedding corresponding of x_k. As a result, LTSA embeddings do not preserve the aspect ratio of the data manifold. In contrast, we use rigid transformations for alignment with no assumption on Y, preserving the aspect ratio and thus incurring less distortion.

<u>Tearing data manifolds:</u>

To embed a data manifold that is either closed, non-orientable or exhibits high curvature, into the intrinsic dimension, the manifold must be torn apart (as a cartographer flattens the three-dimensional earth into a two-dimensional map). We achieve this in the following iterative procedure. At every iteration, we first compute local neighborhoods $(\widetilde{\mathcal{N}}_k)_1^n$ in the embedding. Then the i*i*th intermediate neighborhood $\widetilde{\mathcal{V}}_i$ in the embedding is the union of the neighbors of the points in cluster $\mathcal{C}_i$, *i.e.* $\cup_{x_k \in \mathcal{C}_i} \widetilde{\mathcal{N}}_k$. Note that these may differ from the intermediate neighborhoods in the data $(\mathcal{V}_i)_1^m$. Given these intermediate neighborhoods in the data and the embedding, instead of aligning all overlapping intermediate views we align *only* those views that are overlapping in both the data ($\mathcal{V}_i \cap \mathcal{V}_j \neq \emptyset$) and the embedding ($\widetilde{\mathcal{V}}_i \cap \widetilde{\mathcal{V}}_j \neq \emptyset$) *i.e.*, we solve for orthogonal matrices $(S_i)_1^m$ and translation vectors $(t_i)_1^m$, that minimizes the alignment error,

$$\sum_{k: x_k \in (\mathcal{V}_i \cap \mathcal{V}_j), y_k \in (\widetilde{\mathcal{V}}_i \cap \widetilde{\mathcal{V}}_j)} \left\| (S_i^T x_{k,i} + t_i) - (S_j^T x_{k,j} + t_j) \right\|_2^2 .$$

In particular, we define a bipartite graph Γ' such that $V(\Gamma') = V(\Gamma)$ while $(k, i) \in E(\Gamma')$ if $x_k \in \mathcal{V}_i$ and $y_k \in \widetilde{\mathcal{V}}_i$. Finally, we solve the Eq. (1) by replacing Γ with Γ' throughout. In each iteration, the alignment results in increased overlap between intermediate neighborhoods in the embedding. At the end of the procedure, for manifolds such as a sphere or a Kléin bottle, a tear is retained, while for others, such as a swiss roll or barbell, the tear vanishes as the intermediate views are fully stitched together.

Gluing instructions for the torn embedding:

To visually represent points $y_l \sim y_{l'}$ that are in close proximity in the data space but appear on opposite sides of the tear, we color their embeddings with a similar color.

We say that the embedded points $y_l \sim y_{l'}$ if the following conditions are satisfied

1. The points x_l and $x_{l'}$ lie in different clusters $c_l \neq c_{l'}$
2. The points x_l and $x_{l'}$ are in the intersection of the corresponding intermediate views in the data i.e. $\mathcal{V}_{c_l} \cap \mathcal{V}_{c_{l'}}$
3. The embeddings y_l and $y_{l'}$ are far apart as captured by the empty intersection of the corresponding views in the embedding $\widetilde{\mathcal{V}}_{c_l} \cap \widetilde{\mathcal{V}}_{c_{l'}}$.

A graph on the embedded points where edges correspond to the relation $y_l \sim y_{l'}$ is constructed. The coloring of the tear is obtained as the eigenvectors of the Laplacian of the above graph. It serves as the "gluing" instructions for the torn embedding to recover the topology of the data manifold (Figure 1(e, f, g, h), Figure 2(b, f), Supplementary figure 5, Supplementary figure 6).

Tear-aware Shortest Path Distance:

Let $d_0(y_k, y_k)$ be the naive shortest path distance between y_k and $y_{k'}$ computed using Dijkstra algorithm on the k-nearest neighbor graph on the embedding (k=10 in all of our experiments). If $y_l \sim y_{l'}$ then, by definition, the two points have local representations due to both the c_lth view (x_{l,c_l} and x_{l',c_l}) and the $c_{l'}$th view ($x_{l,c_{l'}}$ and $x_{l',c_{l'}}$). Then we define the t-hop shortest path distance $d_t(y_k, y_{k'})$ as the minimum of $d_{t-1}(y_k, y_{k'})$ and

$$\min_{y_l \sim y_{l'}} d_{t-1}(y_k, y_l) + \min\{\|x_{l,c_l} - x_{l',c_l}\|_2, \left\|x_{l,c_{l'}} - x_{l',c_{l'}}\right\|_2\} + d_{t-1}(y_{l'}, y_k).$$

The tear-aware shortest path distance $\widetilde{d}(y_k, y_{k'})$ is obtained by computing d_t for increasing values of $t \in \{1, \ldots, t_{max}\}$ until the convergence of d_t. We use $t_{max} = 5$ in all of our experiments.

Global distortion calculation:

Let l_{kj} and $\widetilde{l}_{kj}$ be the shortest path distances from x_k to x_j and from y_k to y_j, respectively. These are computed using the Dijkstra algorithm on the k-nearest neighbor graph on the ground truth data and the embedding. Note that for the torn embedding we instead use the tear-aware shortest path distances as defined above. Then the global distortion at the kth point is given by

$$\mathbb{G}_k = \max_{j \neq k} \frac{\widetilde{l}_{kj}}{l_{kj}} \max_{j \neq k} \frac{l_{kj}}{\widetilde{l}_{kj}}.$$

Finally, to measure the quality of a global embedding , we define its distortion to be a histogram built from global distortions at every point *i.e.* $(\mathbb{G}_k)_1^n$.

<u>Construction of k-nearest neighbor graph on the data and the embedding:</u>

To calculate shortest path distance for the global distortion calculation, we calculate k-nearest neighbor graphs on both the data and the embedding. The number of nearest neighbors, *k*, employed to construct the k-nearest neighbor graph is determined as the minimum value within the range of [9, 25] at which the distances of shortest paths on the data exhibit convergence. To ascertain this convergence, the following procedure is undertaken: The mean absolute relative change in the shortest path distances is computed for consecutive values of *k*. These are subjected to a threshold of 1e-3. The smallest value is identified for which the relative change remains below the tolerance for three consecutive values. In the rare situation where none of the value satisfies this condition, we relax the tolerance by doubling it, and repeat the process.

<u>Hyperparameter optimization for manifold learning techniques</u> (UMAP, ISOMAP, t-SNE, LTSA, LDLE, RATS). We performed a grid search optimization based on the principal hyperparameters of each method. The optimal embedding was based on the hyperparameter set that yielded the minimal value for the maximum of the pointwise global distortions i.e. $\max_k \mathbb{G}_k$. The following table shows the ranges of hyperparameters tested during the grid search optimization for each of the competing methods:

Method	Hyperparameters (all pairs are tested)
UMAP	`n_neighbors` in {10, 25, 50, 100, 200, 300} `min_dist` in {0.1, 0.25, 0.5, 0.75, 1}
t-SNE	`perplexity` in {25, 50, 100, 200, 300, 500} `early_exaggeration` in {8, 16, 32, 64, 128, 256}

ISOMAP	`n_neighbors` in {5, 10, 15, 20}
LTSA	`n_neighbors` in {5, 10, 25, 50, 100, 200, 300}
LDLE	`k` in {28, 21, 14}, `eta_min` in {5, 10}, `gl_type` in {"unnorm", "diffusion"}
RATS	`k` in {28, 21, 14}, `eta_min` in {5, 10}

Table 1. Hyperparameter range of the different methods

Kruskal-Wallis Statistics

Given multiple groups (manifold learning algorithms) with non-normal distribution statistics, we administer a non-parametric Kruskal-Wallis test throughout to consider the null hypothesis that the pointwise global distortion due to each algorithm (groups) originates from the same distribution. The test is implemented using the scipy.stats.kruskal method from the scipy Python package with default hyperparameters to obtain the H statistic and the p-value. At a significance criterion of 1e-3, the null hypothesis is rejected across all the datasets considered in this work.

Two-sided Wilcoxon Test Statistic

To test the null hypothesis that the pointwise global distortion due to the embeddings generated by two algorithms originates from the same distribution, we employ the two-sided Wilcoxon signed-rank test. The test is implemented using scipy.stats.wilcoxon method from the scipy Python package with default hyperparameters to obtain the W statistic and the p-value, and a significance level of 1e-3 is used throughout.

Grid cell population datasets:

We took the simultaneously recorded extracellular spikes from the 149 “pure” grid cells identified by [13] from the MECII region of a rat's brain while it forages in an open field . Similar to their preprocessing procedure for the topological analysis of this data, we preprocessed the neural activity by first applying Gaussian smoothing with a window size of 50ms, then selecting the 15000 most active population vectors which is then followed by a PCA projection into 6 dimensions. We follow the downsampling technique used in[13], which is based on a point-cloud density strategy motivated by a topological denoising technique[48] and a fuzzy topological representation used in UMAP[22]. For downsampling we use a cosine metric and a neighborhood size of 1500, and obtain a topologically denoised subsample of size

4800. Finally, we embedded this 6-dimensional dataset using various manifold learning techniques. To be consistent with the analysis in the original study[13], we use the cosine metric with all techniques and specifically, we use kernel PCA with the cosine kernel to construct local views with RATS. Additionally, while LTSA is limited to linear-local views of the data, RATS adopts a more flexible approach, allowing the utilization of kernel techniques for modeling the data locally.

The rough-edges that sometimes appear along the torn RATS embeddings for closed manifolds can be mitigated by adjusting the tear-positioning using a semi-automated cut-and-paste method within a graphical user interface (Supplementary Figure 6). This process maintains the same topology while producing an embedding with a smoother tear. The firing rate maps of the grid cells are visualized in spatial coordinates and on the smoothed RATS embedding (Supplementary Figure 7).

Similar to our distortion calculations, for the persistent homology analysis, we also constructed *k*-nearest neighbor graphs. Using these, we computed the shortest-path distances between all pairs of points in the data and their embeddings. For the RATS and LDLE embeddings, we utilized tear-aware shortest path distances between the subset of points that lie on the tear. Finally, we calculated the persistence barcodes for the zeroth and first cohomology by applying RIPSER[49] to the shortest path distance matrices (Supplementary Figure 8).

<u>Head direction cell population dataset</u>:

We took the spiking activity of 22 neurons in the anterodorsal thalamic nucleus (ADn) region of a mouse that was awake and foraging in an open environment along variable paths with variable velocities [12]. The spike activity is preprocessed in exactly the same manner as in [12]: firing rates are first computed by convolving the spike times with a Gaussian kernel of standard deviation 100 ms. Then the rates are replaced by their square root to stabilize the variance. Finally, the first 5000 rates form the dataset in our analysis. We embedded this 22-dimensional dataset using various manifold learning techniques.

<u>Network training, tasks, and manifold generation:</u>

Idealized datasets provide validation of ground-truth manifolds, whereas in real-world datasets, often the ground-truth is unknown. Under these circumstances, neural network models can provide a means to generate large-scale datasets whose underpinnings obey manifold-like topologies, which are also riddled with realistic features, such as, uneven density, noise, discontinuities. In this work we also

develop methods to control and manipulate noise along and off the manifold, as well as vary the density of data points across the manifold. This allows us to investigate the impact of these features on distortion using different techniques.

Our examples are drawn from the field of neuroscience, where Recurrent Neural network models (RNNs) have demonstrated remarkable success in recapitulating neural population activity[50–53]. Specifically, we utilize topologies generated by the dynamics of neural networks representing three neuroscientific behavioral tasks [33–35]. Each of these networks emulates the properties and encoding principles of neural population dynamics believed to underlie various behaviors in the macaque cerebral cortex and the mouse cerebellum. By employing RNNs, we not only recreate the dynamics underlying these tasks but discover families of population trajectories that remain confined to manifold-like geometries with quantifiable intrinsic dimensions.

Motor Timing network datasets: In the originally reported motor timing task [33], monkeys were trained to respond to color cues (Cue) to generate time intervals of different durations. Figure 2a illustrates the task structure, where after the presentation of the Cue, a transient Start flash is presented (Start). The monkey was trained to generate different time intervals by a button press or a saccadic eye movement (Response). This led to the generation of response intervals (Tp) from the onset of the Start flash that depended on the cued context. The error feedback with respect to the desired time intervals (Ts) was provided by enforcing a time window within which responses were rewarded. The original work[33] also put forth a hypothesis that the contextual signal, encoding context-dependent timing, was supplied to the cerebral cortex via step-like inputs of graded magnitudes. Moreover, this signal interacts with recurrent cortical architectures to generate flexible timing signals. Following these findings, the RNN implementation assumes that the recurrent connectivity models the cortical region responsible for encoding time production signals. Additionally, a time-dependent step input is used to model the subcortical contextual cue, whose magnitude sets the time interval production dynamics in the cortex.

The RNN consisted of $N = 200$ nonlinear units with a hyperbolic tangent activation function. Let the activity of the population be represented by the time-varying vector $\mathbf{x}(t)$, whereas the firing rate of such a population be vector $\mathbf{r}(t)$. We assume that the time constant of each unit τ is 10 ms. Each unit also contains a bias term (c_x) and a term for internal noise (drawn from $\mathcal{N}(0, 0.01)$), which was sampled at each time step (δt = 1 ms). The Start pulse is administered at a random time after the presentation of

the contextual Cue input to the network. The target function was a ramp specified after the Start until the network output crossed a threshold at the desired time interval for each trial t_s. The actual time at which the network crossed the threshold (Response) was taken as production time for a given trial t_p.

The network dynamics were governed as per the following equations:

$$\tau \dot{\boldsymbol{x}}_{\boldsymbol{i}}(t) = \boldsymbol{x}_{\boldsymbol{i}}(t) + \sum J \boldsymbol{r}_{\boldsymbol{i}}(t) + \sum B \boldsymbol{u}(t) + \boldsymbol{c}_x + \rho(t)$$

$$z(t) = \sum_{j=1}^{N} w_{oj} \mathbf{r}_j(t) + c_z$$

The recurrent weight connectivity matrix J was initialized by drawing values from a normal distribution with zero mean and variance $\frac{1}{N}$[54]. The network received a two-dimensional input $\mathbf{u}(t)$ consisting of a context cue $u_c(t)$ and a transient pulse at a later time, $u_s(t)$, which marked the time from which the network generated the desired time interval. Model parameters, J, B, w_o, C, were trained using backpropagation through time[55] by minimizing a squared loss function between the network output $z(t)$ and a target function, which was a linear ramp starting at the time of Start and crossing a threshold at the desired time T_s for a given context input.

Trace conditioning network datasets: In the second task of *Trace conditioning*, mice learned to associate between two sensory cues, viz., a brief flash of light (Light) and a brief periocular stimulation (Airpuff) after a fixed delayed interval. This behavior is thought to be encoded in cerebellar cortical circuitry via basis set inputs from cerebellar granule cells. The second source of the input signaling sensory events arrives at the cerebellar cortex via the climbing fibers. To mimic this input regime carrying sensory and temporal information, RNNs received multi-dimensional time-varying input $u(t)$ through synaptic weights $B = [b_0, b_g c, b_c f]$. The input $u(t)$ included a step input, b_0, the summation of decaying Gaussian granule cell basis kernels $b_g c$ for the Light and Airpuff, and a 10 ms pulse to mark climbing fiber activation, $b_c f$. The network was trained to cross a threshold in anticipation of the time of the airpuff, however, no explicit form of the target function was imposed on the network. The target function was only defined at brief time points (10 ms) on three occasions, viz., at the mean of the distribution on the threshold, at the start of the trial at zero, and at the end of the trial at zero again. Despite the varied shapes of the training output solution by various networks, it was observed that expert networks followed the statistics of the stimulus for both paired and test trials. However, the statistics of the output

of the switch networks were not different for the distribution conditions for paired and test trials, although their unit activities exhibited significant heterogeneity that matched that observed in neural responses. Network training was carried out in a manner similar to previous work[14,33].

Interval reproduction network datasets: In the third task, participants measured a sample interval and reproduced it immediately after measurement, similar to previous work[56]. On each trial, these intervals were sampled from a uniform prior distribution of time intervals (Prior). Previous work[34,56,57] demonstrated that monkey and human timing responses during this task were biased towards the mean of this prior distribution in a manner predicted by Bayesian estimators (such as Bayes Least Squares or BLS). In the absence of prior knowledge, one would expect no bias towards the mean of the prior but an increase in variance, mimicking scalar variability in timing, as predicted by non-Bayesian models like Maximum Likelihood estimation (MLE)[37]. When neural activity was recorded in the macaque dorsomedial frontal cortex during this task, population dynamics assumed a semicircular trajectory, which when read out linearly gave rise to states that were compressed around the mean of the prior interval. These compressed neural states could account for both the Bayesian model predictions as well as the monkey's responses, which were also consistent with Bayesian models. This leads to the hypothesis that the curvature of the semicircular trajectory was encoding prior knowledge, leading to compressed states when read out linearly. However, if the trajectory is unfolded this leads to the converse hypothesis that if instead the state estimates were read along the manifold geodesic, the outcomes would represent responses incurred in the absence or after the erasure of prior knowledge, consistent with Maximum Likelihood estimation (MLE).

We use the RNN to recreate this task and the underlying neural dynamics during the measurement epoch leading to the construction of a three-dimensional manifold, where the third dimension is orthogonal to the task-related manifold. Based on earlier work, it was expected that linearly decoding the curved manifold (by projecting onto its diameter) in its ambient dimension of three would yield readouts consistent with Bayesian models. On the other hand, embedding this manifold into two dimensions using manifold learning techniques and decoding along geodesic distance was expected to yield readouts consistent with MLE.

To achieve this, RNNs were trained using a two-dimensional time-varying input $u(t)$ that consisted of a constant drive whose offset on each trial was sampled from a Gaussian distribution ($\mathcal{N}(0.5, 0.1)$), a brief pulse input (20 ms) that demarcated the Start of the trial, after which a time interval was sampled from a discrete uniform distribution (529-1059 ms, comprising ten intervals), which was followed by another 20

ms pulse to mark the end of the time interval. The network was trained to cross an arbitrary threshold ($z(t) = 1$) at the time of the sampled interval. During training, duration-dependent noise (Weber fraction dependent = 0.2) was introduced into the time of the target. Training under such scalar noise conditions induces the behavioral output to follow a compression towards the mean of the prior, which is the statistically optimal solution for such conditions.