COMET: COMMUTE-ORDER MAPPING OF EXPECTED TOPOLOGY FOR FAITHFUL DATA VISUALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visualization methods such as t-SNE, UMAP and their successors are the standard tools for making high-dimensional data visually interpretable. However, they expose hyperparameters that substantially change the embedding, with no ground truth to guide the choice. The settings are often chosen by visual inspection instead, which may bias the result toward the user's prior expectations. The embeddings also tend to break continuous structure into pieces and to distort both scale and global geometry. We introduce COMET (Commute-Order Mapping of Expected Topology), which embeds the geometry of a neighborhood graph. Commute times on that graph supply distances that follow the manifold, and ranking places them on a common scale with the graph's local similarities. Every neighbor and sampled non-neighbor then receives a target probability in the cross-entropy objective that calibrates the repulsive terms alongside the attractive ones, rather than every non-neighbor sharing the same target probability as in UMAP. COMET scales to large datasets and requires no hyperparameter tuning. Against seven established methods on four synthetic datasets with known topology and four real datasets, it ranks first on 21 of 24 comparisons under three validation measures built to quantify geodesic distance preservation, fragmentation and scale distortion.

1 INTRODUCTION

Dimensionality reduction is routinely used to make high-dimensional data visually interpretable by embedding it in two or three dimensions. However, verifying that the embedding is faithful would require the very structure it was created to reveal, so in general there is no ground truth to check it against. Widely used methods nonetheless expose tunable hyperparameters that can change the embedding substantially — t-SNE's perplexity (van der Maaten & Hinton, 2008), UMAP's minimum separation (McInnes et al., 2020), TriMAP's triplet weighting (Amid & Warmuth, 2019) — with no mechanism for determining which setting is correct for a given dataset (Wang et al., 2021). In practice these parameters are often tuned by eye (Wattenberg et al., 2016), which leaves the user's prior expectations to decide when the result is acceptable. The embedding then confirms the structure it was tuned to display rather than providing evidence for it (Staff, 2025). Tuning against a validation measure only relocates the problem, biasing the embedding toward whatever that measure rewards at the expense of the properties it does not capture (Venna et al., 2010). The lack of any principled basis for setting these parameters is only half of the problem. Existing methods also tend to fragment continuous structure into disconnected pieces and to distort both scale and global geometry. (Wattenberg et al., 2016; Wang et al., 2021; Kobak & Linderman, 2019).

The two most commonly used methods constrain geometry only within a local scale defined by a single parameter. t-SNE (van der Maaten & Hinton, 2008) assigns pairwise similarities using a Gaussian kernel whose width is fixed by the perplexity, a point well beyond that width receives a similarity numerically indistinguishable from zero. UMAP (McInnes et al., 2020) builds a neighborhood graph and assigns every sampled non-neighbor a similarity of zero. Neither objective distinguishes a point beyond its local scale from one on the far side of the manifold, and what global structure their embeddings display is largely inherited from the initialization (Kobak & Linderman, 2019).

More recent methods such as PaCMAP (Wang et al., 2021), TriMAP (Amid & Warmuth, 2019) and LocalMAP (Wang et al., 2025) do constrain structure beyond the local scale, through mid-near pairs or triplets that rank one point nearer to an anchor than another. They differ in how these are sampled

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and weighted, but each measures distances in the ambient space, which can cut across a curved manifold rather than following it (Tenenbaum et al., 2000). PHATE (Moon et al., 2019) and SASNE (Wängberg et al., 2022) avoid this, deriving distances from the neighborhood graph rather than from the ambient space. However, while SASNE has no tunable parameters, its all-pairs computation makes it scale poorly to large datasets. PHATE scales to large datasets only through landmarking, embedding a fixed-size subsample and extending the result to the remaining points. Its diffusion time scale is selected automatically, but its kernel still exposes a tunable decay exponent.

Commute-Order Mapping of Expected Topology (COMET) supplies the missing global geometry, scales to large datasets without landmarking, and requires no hyperparameter tuning. Commute times on a sparse neighborhood graph provide distances that follow the manifold, and a truncated spectral decomposition of the graph Laplacian approximates them efficiently. Ranking those distances puts the graph's local similarities and its global geometry on a common scale. The final cross-entropy objective is calibrated in both its attractive and repulsive terms, with a target probability set by rank for both neighbors and sampled non-neighbors. Rather than treating kernel bandwidths, repulsion strengths and regularization terms as tunable hyperparameters, COMET derives them from the data.

Trustworthiness and Continuity (Venna & Kaski, 2006), the standard local validation measures, rank local neighborhoods by ambient distance. Neither registers a continuous manifold torn into pieces, one region compressed relative to another, or distance ranks scrambled beyond the local scale. Section 3.2 adds three measures for what they miss: geodesic distance preservation, fragmentation and scale distortion. Against seven established methods on four synthetic datasets and four real datasets, a fixed configuration of COMET ranks first on 21 of the 24 comparisons they define.

2 The COMET Algorithm

2.1 The Graph as a Discrete Manifold Approximation

Under the manifold hypothesis (Fefferman et al., 2016), high-dimensional data lies on a lower-dimensional manifold. Because this manifold is curved in general, ambient distances between distant points can cut across ambient space, short-circuiting the manifold geometry (Tenenbaum et al., 2000). To discretely approximate the manifold, COMET constructs a k-nearest neighbor (k-NN) graph, restricting ambient distance evaluations to locally Euclidean regions.

Given a dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and a dissimilarity measure δ, an undirected graph $G = (V, E)$ is constructed with nodes $V = \{1, \ldots, n\}$, where an edge $\{i, j\} \in E$ exists if x_j is among the k nearest neighbors of x_i, or vice versa. Let $\mathcal{K}_i$ denote the k nearest neighbors of node i, and $\mathcal{N}_i = \{j \mid \{i, j\} \in E\}$ its complete neighbor set, so that $\mathcal{K}_i \subseteq \mathcal{N}_i$ with $|\mathcal{K}_i| = k$ for every i. Since the spectral construction of Section 2.2 requires a connected graph, any disconnected components of G are bridged by their minimum spanning tree (Appendix C.2).

Converting dissimilarities to edge weights requires a kernel and a scale. A fixed global scale divides every dissimilarity by the same value, so the weights saturate where sampling is dense, leaving neighbors indistinguishable, and vanish where it is sparse, leaving nodes weakly connected. COMET therefore uses an exponential kernel whose scale is set per node. For each edge $\{i, j\} \in E$ two directed weights are computed, the one from i to j being:

$$w_{i \to j} = \exp\left[-\frac{\delta_{ij} - \delta_{\min}(i)}{\overline{\delta}_i - \delta_{\min}(i)}\right], \tag{1}$$

where $\delta_{ij} = \delta(x_i, x_j)$, $\overline{\delta}_i = \frac{1}{|\mathcal{N}_i|} \sum_{v \in \mathcal{N}_i} \delta_{iv}$ and $\delta_{\min}(i) = \min\{\delta_{iv} \mid v \in \mathcal{N}_i\}$.[1]

The rapid decay of the exponential kernel concentrates weight on the most immediate neighbors, where the locally Euclidean approximation is most secure and the risk of short-circuiting the manifold lowest. Neither the shift $\delta_{\min}(i)$ nor the scale $\overline{\delta}_i - \delta_{\min}(i)$ in equation 1 is chosen. Both are derived to maximize the differential entropy of the weights produced by $f_{i \to j}(a, \tau) = \exp\left[-(\delta_{ij} - a)/\tau\right]$, subject to weights in $(0, 1]$ (Appendix C.1). Maximizing entropy spreads the weights across $(0, 1]$, so neighbors at different distances receive distinguishable weights rather than saturating at either end.

[1] If all nodes connected to i are equidistant from it, as can happen with some discrete dissimilarity measures, both the scale and every numerator vanish and the weights are set to 1.

The directed similarities of equation 1 are symmetrized by their geometric mean, $w_{ij} = \sqrt{w_{i\to j} w_{j\to i}}$, yielding the undirected weighted graph $\bar{G} = (V, E, w)$. Compared to additive symmetrization schemes, such as the arithmetic mean or the fuzzy union used in UMAP (McInnes et al., 2020), which retain large weights even when a connection is unreciprocated, the geometric mean penalizes one-directional connections. This reduces the influence of hub nodes that can emerge as artifacts in high-dimensional k-NN graphs (Radovanovi et al., 2010).

2.2 Manifold-Aware Distances via Commute Times

While the edge weights w_{ij} resolve local geometry, capturing the global structure of the manifold requires a robust graph distance. Unlike the shortest-path distance, which is sensitive to noise (Dijkstra, 1959), or the diffusion distance, which relies on a user-defined time-scale parameter (Coifman & Lafon, 2006), the commute-time distance (CTD) (Fouss et al., 2007) integrates over all pathways, making it robust to noise, and introduces no tunable parameters. By representing the expected round-trip time of a random walker between any two nodes, it grows when the connection between them rests on a few tenuous paths and shrinks when many paths connect them.

Let W be the $n \times n$ symmetric weight matrix of $\bar{G}$, and D the diagonal degree matrix with $D_{ii} = \sum_v w_{iv}$. The structural properties of the graph are accessed via its normalized symmetric Laplacian (von Luxburg, 2007): $L_{sym} = I - D^{-1/2} W D^{-1/2}$. Excluding the trivial eigenvalue $\lambda_0 = 0$, let $0 < \lambda_1 \leq \cdots \leq \lambda_{n-1}$ denote the non-trivial eigenvalues of L_{sym}, with eigenvectors $u_1, \ldots, u_{n-1}$. Ordered by increasing eigenvalue, these eigenvectors constitute the graph analogue of a Fourier basis: the smallest eigenvalues correspond to modes that vary smoothly across the graph, encoding its global structure, while the largest correspond to modes oscillating between nearby nodes, encoding local structure and noise. Computing the exact CTD requires the entire spectrum of L_{sym}, which is computationally infeasible for large datasets. COMET therefore truncates the spectrum to the l smallest non-trivial eigenvalues, keeping the low-frequency modes that carry global structure and discarding the high-frequency modes that carry local structure. A regularization term $\epsilon \geq 0$ is also added to each retained eigenvalue to bound the condition number of L_{sym} when the graph is weakly connected (Appendix D.1). The resulting approximate CTD between nodes i and j is given by[2]

$$\tilde{c}(i,j) = \sum_{m=1}^{l} \frac{1}{\lambda_m + \epsilon} \left(\frac{u_{im}}{\sqrt{D_{ii}}} - \frac{u_{jm}}{\sqrt{D_{jj}}} \right)^2, \tag{2}$$

where u_{im} is the ith entry of u_m. Since each summand is a squared coordinate difference the approximate CTD between nodes i and j can be expressed as the squared Euclidean distance $\tilde{c}(i,j) = \|(C_l)_i - (C_l)_j\|^2$ between rows i and j of the approximate commute-time embedding

$$C_l = D^{-1/2} U_l (\Lambda_l + \epsilon I)^{-1/2}, \tag{3}$$

where Λ_l is the diagonal $l \times l$ matrix of the retained eigenvalues and U_l the corresponding $n \times l$ matrix of eigenvectors. The criterion for determining the truncation parameter l is detailed in Section 2.3.

2.3 Placing Local and Global Structure on a Common Scale

The similarities w_{ij} are defined only on the edges of $\overline{G}$ and encode local structure, while $\tilde{c}$, with its high-frequency modes discarded, captures only global structure. Both are needed, on a comparable scale that a similarity and a distance do not share. Rank provides that scale by using only the ordering of each quantity. Within the neighborhood of node i, sorting $\mathcal{K}_i$ in decreasing order of w_{ij} yields the sequence of nodes $j_1, \ldots, j_k$, with ranks $1, \ldots, k$. Beyond this neighborhood, where ambient distances are no longer reliable, the ordering is instead supplied by the approximate CTD $\tilde{c}(i, \cdot)$.

Ranking all $n - k - 1$ remaining points for every node is $O(n^2)$. Consequently, COMET samples non-neighbors and assigns expected population ranks according to their positions in the sampled ordering. For each node i, s points are drawn uniformly at random without replacement from the $n - k - 1$ nodes outside $\mathcal{K}_i \cup \{i\}$. Let $\{j_{k+1}, \ldots, j_{k+s}\}$ denote this sample, indexed so that $\tilde{c}(i, j_{k+1}) \leq \cdots \leq \tilde{c}(i, j_{k+s})$.

[2]The constant factor $\mathrm{Vol}(\overline{G}) = \sum_i D_{ii}$ is omitted, as the embedding is rescaled in subsequent steps. Moreover, since $\overline{G}$ is guaranteed to be connected, $\lambda_1 > 0$, making equation 2 well defined even when $\epsilon = 0$.

Uniform sampling makes estimation of the expected rank a classical problem in order statistics (Arnold et al., 1992, §3.7). Shifting by the k ranks reserved for the neighborhood, the expected rank of the t-th sampled point j_{k+t} is given by

$$r_t = k + \frac{t\,(n-k)}{s+1}, \qquad t = 1, \ldots, s. \tag{4}$$

These ranks satisfy $k < r_1 < \cdots < r_s < n$ and divide the interval $[k, n]$ into $s+1$ equal parts, as illustrated on the horizontal axis of Figure 1. Combining the neighbors with the s sampled non-neighbors then yields a set of target nodes $J_i = \{j_1, \ldots, j_{k+s}\}$ for each anchor node i, with ranks $R = \{1, \ldots, k, r_1, \ldots, r_s\}$, the same for all i, encoding both local and global manifold geometry.

The sample size s sets the resolution of the nonlocal ordering, with larger values providing more finely spaced expected ranks. The spectral dimension l of Section 2.2 is matched to this resolution by setting $l = s$. This coupling is motivated by the geometry around each anchor: the s anchor–target difference vectors span at most s dimensions. Consequently, we match the spectral dimension l to the maximum dimension needed to represent the s difference vectors of a single anchor. While this does not guarantee that the subspace preserves the exact geometry around every anchor, COMET's requirement is much weaker: preservation of anchor–target distance ranks rather than exact geometry.

2.4 From Ranks to Neighbor Probabilities

A position in an ordering is not yet a measure of how close two points should be in the embedding. Target node ranks are therefore mapped to target probabilities by a monotonically decreasing function:

$$p(r) = \left(1 + \frac{r}{2\sigma}\right)^{-2}, \qquad r \in R, \tag{5}$$

where σ sets the rate of decay. Because the ranks in R range from 1 to nearly n, the kernel must decay rapidly across the first k, resolving local structure by separating a point's nearest neighbors, and slowly beyond them, preserving the global structure carried by the difference between the hundredth and the thousandth nearest node. The exponential kernel $\exp[-r/\sigma]$ meets the first requirement but not the second, compressing the probabilities of the sampled ranks toward zero, while the Gaussian kernel $\exp[-r^2/2\sigma^2]$ meets neither, being flat about the origin and decaying even faster than the exponential. The rational kernel in Equation 5 satisfies both as illustrated in Figure 1.

The two halves of R scale differently with n: the neighborhood ranks remain $1, \ldots, k$ while the sampled ranks grow with the dataset. Under a fixed σ the sampled target probabilities drift toward zero as n grows, until they are indistinguishable from one another. COMET therefore sets $\sigma = r_1$, the first sampled rank, which pins $p(r_1) = 4/9$ at every n. The k neighborhood probabilities then approach one as n grows, as they should, since a fixed k is a smaller fraction of a larger dataset.

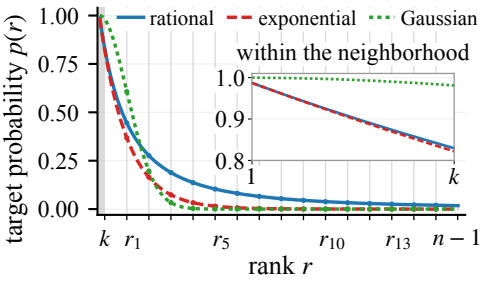


Figure 1: **Mapping ranks to target probabilities.** The k ranks read from the similarities (shaded) and the estimated ranks $r_1, \ldots, r_s$ of equation 4 on a common axis, each mapped to a target probability by equation 5, with the exponential and Gaussian kernels at the same scale ($\sigma = r_1$) for comparison. Shown for $k = s = 15$ and $n = 1000$. *Inset:* behavior within the k-nearest neighborhood.

2.5 A Fully Calibrated Cross-Entropy Objective

COMET produces a low-dimensional embedding $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}^m$, where m is the dimension (typically $m \in \{2, 3\}$), in which similarity between points i and j is measured by the heavy-tailed Cauchy kernel (van der Maaten & Hinton, 2008): $q_{ij} = \left(1 + ||y_i - y_j||^2\right)^{-1}$. Its tail alleviates crowding by allowing moderate distances in d dimensions to be represented by much larger ones in m, mitigating the reduction in spatial volume.

The commute-time embedding of equation 3 supplies the starting point for the embedding, and its coordinates carry an arbitrary magnitude. It is therefore expressed in units of a typical graph edge, $\widetilde{C}_l = C_l \,/\, \mathrm{median}_{\{i,j\}\in E} \|(C_l)_i - (C_l)_j\|$, and the embedding is initialized from the first m columns

of $\widetilde{C}_l$. Without this rescaling the initial distances bear no relation to the unit scale of q, and gradient descent can begin with every pair at $q \approx 0$ or $q \approx 1$, where the gradient vanishes.[3]

Since the rank assigned to the t-th target node is the same for every anchor, namely $R_t = t$ for the k neighbors and $R_t = r_{t-k}$ for the s sampled points beyond them, applying equation 5 yields a single target probability vector shared by all nodes: $p_t = p(R_t)$ for $t = 1, \ldots, k+s$. Writing j_t for the t-th target node of anchor i, the embedding is optimized by minimizing the binary cross-entropy between these target probabilities and the embedded similarities, evaluated over the $k+s$ target nodes of each anchor:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{t=1}^{k+s}\left[p_t \log q_{ij_t} + (1-p_t)\log(1-q_{ij_t})\right]. \tag{6}$$

In UMAP's negative sampling every sampled negative pair shares one target probability (McInnes et al., 2020), so repulsion carries no information about how far apart two points should be. In equation 6 both the attractive term $p_t \log q_{ij_t}$ and the repulsive term $(1-p_t)\log(1-q_{ij_t})$ are calibrated by the rank (or expected rank) of the target node j_t relative to the anchor i. The sampled nodes are periodically redrawn during training to broaden coverage of the $n-k-1$ candidates. A fresh draw is assigned the same target probabilities p_t, since equation 4 follows from the sampling scheme alone, and ordering it only requires distances in the precomputed commute-time embedding $\widetilde{C}_l$. The objective is minimized by asynchronous stochastic gradient descent.

2.6 Computational Complexity

Every stage of COMET is linear in n for fixed k and s, apart from the k-nearest neighbor search, which carries an empirical cost of $O(n^{1.14})$ when using the NN–descent algorithm (Dong et al., 2011). The truncated eigensolver returns only the $l = s$ smallest eigenpairs of the sparse L_{sym} via Lanczos iteration (Lehoucq et al., 1998), and each epoch evaluates $k+s$ interactions per anchor, so both steps scale linearly in n. Measured runtimes on nested subsamples confirm this scaling, with COMET reaching $n = 86{,}024$ in around twenty seconds (Figure 6). Fitted over the five largest of these samples, where fixed startup cost no longer dominates, runtime grows as $n^{0.98}$ for COMET and as $n^{1.07}$, $n^{1.09}$ and $n^{1.09}$ for PaCMAP, LocalMAP and TriMAP, respectively.

3 Experiments

3.1 Experimental Setup

COMET is compared with seven established methods: t-SNE, UMAP, PaCMAP, LocalMAP, TriMAP, SASNE and PHATE. Every method is evaluated at its library default configuration on every dataset. Tuning any of them per dataset would bias the results, either toward user expectations or toward a particular validation measure. For COMET, a single default configuration is fixed across all experiments: the neighborhood size is set to $k = 15$, matching the defaults of UMAP (McInnes et al., 2020) and PaCMAP (Wang et al., 2021). The non-neighbor sample size is set to $s = 15$, which by design fixes the spectral dimensionality to $l = 15$. This choice is primarily a cost decision: larger s provide no consistent improvement in structure preservation, while runtime at $n = 86{,}024$ rises from 22 s at $s = 15$ to 336 s at $s = 100$ (Appendix B.4). Finally, the Euclidean distance is used as the ambient distance measure on every dataset, for all methods and validation measures in this study.

3.2 Manifold-Aware Validation Measures

Each method's low-dimensional (LD) embedding is compared against the high-dimensional (HD) reference by five complementary measures, specified in full in Appendix A.

Geodesic Rank Correlation (ρ_{Geo}). Whether an embedding preserves the distances on the manifold is assessed first. Geodesics are approximated by shortest-path distances over 15-NN graphs built independently in the HD and LD spaces, and their Spearman rank correlation $\rho_{\mathrm{Geo}} \in [0, 1]$ is reported. When disconnected both graphs are bridged, so every shortest path is finite.

[3] A normalized PCA initialization is used instead when the graph is disconnected at k and requires bridging, when one of the first m axes of $\widetilde{C}_l$ is stretched by a few isolated points, or when $m > l$.

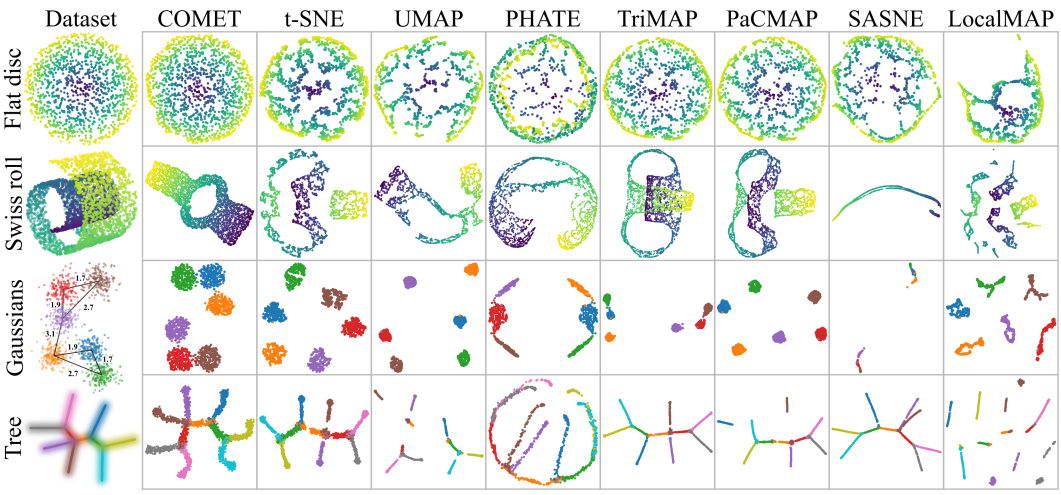


Figure 2: **Synthetic datasets and their embeddings.** *Leftmost column:* the four datasets. The flat disc ($n = 1{,}200$, $d = 3$) is a uniformly sampled 2D disc embedded in 3D. The Swiss roll ($n = 4{,}000$, $d = 3$) is a 2D sheet rolled into 3D with a rectangular hole and Gaussian noise added. The hierarchical Gaussians ($n = 1{,}200$, $d = 3$) form three clusters at each of two levels, annotated with the pairwise separations between centers. The noisy branching tree ($n = 1{,}440$, $d = 60$) has seven branches. *Remaining columns:* 2D embeddings at each method's library defaults.

Manifold Fragmentation (N_{comp}). Geodesic rank correlation can remain high despite fragmentation, motivating a direct connectivity measure. Let $k^\star = \max\{5, k_0\}$, where k_0 is the smallest neighborhood size for which the HD k-NN graph is connected. The floor avoids very small neighborhood sizes, where connectivity can be sensitive to individual neighbor relationships. The number of connected components of the $k^\star$-NN graph built in the LD embedding is denoted $N_{\text{comp}} \in [1, \lfloor n/(k^* + 1) \rfloor]$ [4]. Since $k^\star$ is sufficient to connect the HD data, $N_{\text{comp}} > 1$ indicates that the embedding has fragmented connectivity present in the original data at the same neighborhood resolution.

Scale Distortion (D_{scale}). Rank correlations are invariant to monotone rescaling and Pearson correlation to shifts, so an embedding can score perfectly on either while compressing or inflating distance ratios. Detecting this requires comparing magnitudes across a set of distances that span the data. The Euclidean minimum spanning tree of the HD space serves this purpose: it connects every point through $n - 1$ edges and is unique, fixing the set without a free parameter. Let P be its edge lengths normalized to sum to one, and Q the distances between the same $n - 1$ node pairs measured in the LD embedding, normalized likewise, so each entry is the share of the total length carried by a single edge. Scale distortion is then measured by their Kullback–Leibler divergence, $D_{\text{scale}} = D_{\text{KL}}(P \,||\, Q) \in [0, \infty)$, which is zero when the embedding's distances along these edges are proportional to the data's and grows as they are disproportionately stretched or compressed.

Local Extrusions and Intrusions (Cont, Trust). Local structure preservation is measured by Continuity and Trustworthiness. Continuity penalizes extrusions, where HD neighbors are separated in the LD embedding. Trustworthiness penalizes intrusions, where distant HD points collapse into the same LD neighborhood. Both are computed at $k = 15$, the neighborhood size used throughout, and are scored in $[0, 1]$ with 1 indicating perfect local preservation (Venna & Kaski, 2006).

3.3 Synthetic Data with Known Topology

The methods are first evaluated on four synthetic datasets whose topology is known by construction (Figure 2), each isolating a distinct failure mode. The flat disc contains no structure, so any structure a method produces is fabricated. The Swiss roll adds curvature and a rectangular hole, so a faithful embedding must unroll the sheet while keeping the hole intact. The hierarchical Gaussians of

[4] Every point is joined to its $k^\star$ nearest neighbors, so no component contains fewer than $k^\star + 1$ points. The upper bound is attained when the embedding has isolated clusters of exactly size $k^\star + 1$.

Method	Flat disc			Swiss roll			Hier. Gaussians			Tree		
	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}
COMET	**0.997**	**1.0**	**0.036**	**0.981**	**1.0**	**0.092**	**0.978**	2.2	**0.164**	**0.977**	**1.4**	**0.165**
t-SNE	0.843	2.0	0.071	0.399	24.0	0.378	0.532	6.0	0.357	0.953	6.0	0.417
UMAP	0.862	15.6	0.241	0.825	48.4	0.296	0.929	6.0	0.482	0.795	16.7	0.654
PHATE	0.865	2.8	0.235	0.952	18.0	0.412	0.920	**2.0**	0.311	0.931	4.5	0.323
TriMAP	0.985	**1.0**	0.167	0.201	2.3	0.275	0.974	4.3	0.387	0.976	9.4	0.480
PaCMAP	0.984	2.1	0.179	0.459	3.9	0.410	0.958	6.0	0.355	0.916	12.4	0.591
SASNE	0.951	8.0	0.495	0.939	25.0	0.842	0.974	6.0	1.149	0.973	15.0	0.814
LocalMAP	0.816	6.0	0.322	0.383	26.4	0.672	0.600	7.6	0.835	0.877	20.5	0.866

Table 1: **Validation measures on synthetic datasets, averaged over 10 independent runs.** For $\rho_{\text{Geo}} \leq 1$, higher is better; for $N_{\text{comp}} \geq 1$ and $D_{\text{scale}} \geq 0$, lower is better. Bold values indicate the best performance per column and standard errors are reported in Table 5.

Wängberg et al. (2022) carry their two-level structure in the separations between cluster centers, so a faithful embedding must resolve both the individual clusters and their hierarchical organization. The noisy branching tree of Moon et al. (2019) is continuous throughout, so any disconnection in its branches is an artifact. Table 1 reports the three gloval validation measures, with trustworthiness and continuity in Table 3. The 2D embeddings of all methods are shown in Figure 2.

Flat disc. The established methods fail in two distinct ways on data that contains no structure. Six of the seven fragment the disc, UMAP into 15.6 components on average. TriMAP keeps it intact but distorts its distances, with $D_{\text{scale}} = 0.167$ against COMET's 0.036. Four methods additionally fall below $\rho_{\text{Geo}} = 0.87$, against COMET's 0.997, on data where PCA recovers the disc exactly and attains $\rho_{\text{Geo}} = 1$. COMET returns the disc as a single component with the lowest scale distortion and highest geodesic rank correlation of any method.

Swiss roll. Curvature separates two failure modes that ρ_{Geo} and N_{comp} distinguish between. TriMAP and PaCMAP keep the sheet largely whole, at $N_{\text{comp}} = 2.3$ and 3.9, yet their ρ_{Geo} collapses to 0.201 and 0.459 since they flatten the roll by folding facing layers together rather than unrolling it. The other five methods severely fragment the manifold, into between 18.0 and 48.4 components on average. COMET is the only method that unrolls the sheet as a single component with $\rho_{\text{Geo}} = 0.981$ and the hole preserved (Figure 2).

Hierarchical Gaussians. Five established methods achieve $\rho_{\text{Geo}} \geq 0.92$ yet distort the relative distances between clusters, with D_{scale} between 0.311 and 1.149 against COMET's 0.164. PHATE is the only method to return fewer components than COMET on average, 2.0 against 2.2. However, the remaining validation measures capture what this component count conceals: as Figure 2 shows, PHATE separates the two super-clusters but loses the organization within each and distorts the cluster shapes arbitrarily. COMET's recovery of this structure is not solely inherited from its initialization: replacing its sampled target probabilities with uniform repulsion causes scale distortion and fragmentation under both spectral and PCA initialization (Appendix B.3).

Branching tree. Every established method severs the tree, averaging $N_{\text{comp}} = 4.5$ for PHATE and 20.5 for LocalMAP, against COMET's 1.4. TriMAP attains $\rho_{\text{Geo}} = 0.976$, within a thousandth of COMET's 0.977, while breaking the tree into 9.4 components on average. Several existing methods also distort scale, which D_{scale} registers and Figure 2 shows as a compression of the tree into near one-dimensional filaments rather than the noisy branches COMET recovers.

Trustworthiness and continuity are near-saturated on all four datasets. COMET attains the highest or tied-highest value on both measures for the disc and the Swiss roll, and the highest continuity on the Gaussians. Where it does not lead on both, COMET consistently favors continuity over trustworthiness, accepting local intrusions rather than the extrusions that separate genuine neighbors and fragment the manifold (Venna et al., 2010).

3.4 Real Data Across Four Domains

Four real datasets are evaluated next, with no ground truth available, so the comparison relies on the validation measures. COIL-20 (Nene et al., 1996) comprises $1{,}440$ images of 20 objects photographed

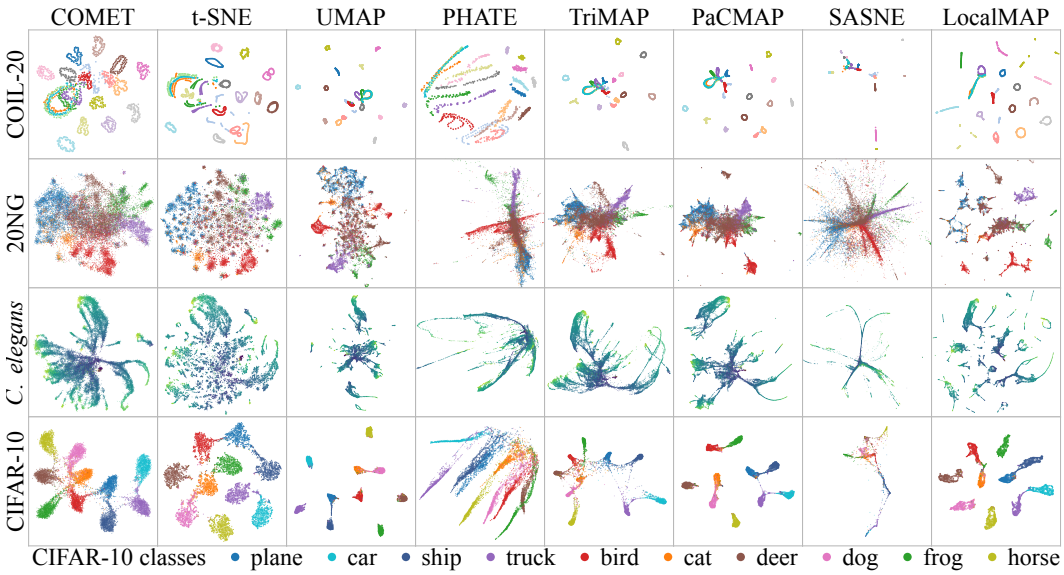


Figure 3: **Embeddings of the real datasets.** 2D embeddings from each method at its library defaults. *Rows:* COIL-20, colored by object identity; 20 Newsgroups, colored by the six top-level topic groups; the *C. elegans* atlas, colored by embryo time; and the CIFAR-10 logits, colored by class. The atlas is subsampled to $n = 20{,}000$ here so that the embeddings of every method can be visualized clearly, while the measures in Table 2 use all $86{,}024$ cells.

at 5° rotation increments, giving twenty closed loops in $16{,}384$ dimensions. The 20 Newsgroups corpus (Mitchell, 1997) consists of $18{,}170$ documents across twenty topics, reduced to $d = 100$ by TF-IDF weighting and truncated SVD. The *C. elegans* embryogenesis atlas of Packer et al. (2019) contributes $86{,}024$ cells, taken in the preprocessed form released by Narayan et al. (2021) as $d = 100$ principal components of the gene expression matrix. Finally, the CIFAR-10 test set (Krizhevsky, 2009) is passed through a ResNet-18 (He et al., 2015), achieving 95.6% test accuracy, and its $d = 10$ output logits are embedded for all $10{,}000$ images. Table 2 reports the global measures and Table 4 trustworthiness and continuity. Figure 3 shows the 2D embeddings for each dataset and method.

COIL-20. On the highest-dimensional data tested, ρ_{Geo} and D_{scale} separate the methods most sharply. COMET reaches $\rho_{\text{Geo}} = 0.708$ against 0.603 for t-SNE, with the remaining six methods below 0.50, and its scale distortion is nearly half the next lowest, $D_{\text{scale}} = 0.263$ against 0.488 for PHATE. It also returns the fewest components, tied with PHATE at 2.2 on average.

20 Newsgroups. Sparse text is the hardest case for every method: no embedding exceeds $\rho_{\text{Geo}} = 0.571$ and all eight lie within 0.17 of one another. COMET is first on geodesic distance preservation and scale distortion, while PHATE returns fewer components, averaging 2.4 against COMET's 3.0.

***C. elegans*.** The atlas is a continuous developmental trajectory, yet 6 of 7 existing methods fragment it, into between 35.6 and 188 components. Only COMET and TriMAP keep it mostly intact at 2.9 and 8.7 components, respectively. Figure 3 shows what those counts correspond to: the trajectories divided into isolated islands rather than the connected branches COMET and TriMAP recover. COMET also attains the lowest D_{scale}, while TriMAP attains the highest ρ_{Geo}, 0.780 against COMET's 0.766.

CIFAR-10 logits. The logits carry a two-level structure of the same kind as the hierarchical Gaussians, ten classes grouped into animals and vehicles. Figure 3 shows that only COMET, TriMAP and SASNE recover it, separating the two groups while resolving the individual classes within each. The remaining five methods resolve the classes but not the grouping. SASNE recovers it at more than three times COMET's scale distortion, $D_{\text{scale}} = 1.112$ against 0.340, and COMET ranks first on all three global measures here.

COMET attains the lowest scale distortion on every dataset, the most consistent result among the three measures. No other method is consistent across the tested datasets. For PaCMAP, TriMAP and LocalMAP, the three methods that use ambient distances beyond the local scale, ρ_{Geo} is highly

Method	COIL-20			20NG			*C. elegans*			CIFAR-10		
	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}
COMET	**0.708**	**2.2**	**0.263**	**0.571**	3.0	**0.462**	0.766	**2.9**	**0.419**	**0.741**	**1.0**	**0.340**
t-SNE	0.603	4.0	0.535	0.514	7.0	1.819	0.607	188.0	1.472	0.504	25.0	1.149
UMAP	0.480	12.2	0.970	0.521	11.6	1.247	0.665	63.1	1.020	0.584	10.9	0.747
PHATE	0.352	**2.2**	0.488	0.484	**2.4**	0.565	0.712	161.4	0.772	0.631	3.4	0.560
TriMAP	0.347	9.8	0.652	0.491	8.1	0.733	**0.780**	8.7	0.573	0.713	2.1	0.366
PaCMAP	0.417	12.7	0.843	0.436	16.9	0.874	0.660	35.6	0.940	0.682	5.4	0.565
SASNE	0.457	11.0	1.743	0.446	6.0	0.630	–	–	–	0.716	3.0	1.112
LocalMAP	0.420	8.0	1.108	0.405	39.5	1.618	0.406	76.1	1.301	0.588	8.1	0.863

Table 2: **Validation measures on the real datasets, averaged over 10 independent runs.** Dashes mark the cells SASNE cannot run due to computational limitations. Bold values mark the best mean per column and standard errors are reported in Table 6. All-pairs shortest paths are infeasible at $n = 86{,}024$, so ρ_{Geo} on the atlas is computed from a single random set of $1{,}000$ landmark nodes.

correlated with how closely those distances approximate geodesics, at rank correlations above 0.85 for all three (Figure 5). On the Swiss roll, COIL-20 and 20 Newsgroups, where that approximation is weakest, their ρ_{Geo} averages 0.35–0.44 against COMET's 0.75. Where it is closest, COMET and TriMAP differ by at most 0.014, with TriMAP fractionally ahead on the atlas dataset. PHATE returns the fewest components on 20 Newsgroups, then the second highest on the atlas, averaging 161.4. COMET ranks first in ten of the twelve columns in Table 2 and second in the other two.

Trustworthiness and continuity follow the pattern of Section 3.3: COMET attains the highest continuity on all four datasets and t-SNE the highest trustworthiness on all four.

4 Discussion and Limitations

The artifacts produced by established visualization methods surveyed in Section 1 — manifold fragmentation, scale distortion and global geometry loss — are not inherent to low-dimensional embeddings. COMET avoids them by combining a local neighborhood kernel with global commute-time ranks, which capture the structure of the data at scales the kernel alone does not reach. Across eight datasets spanning synthetic manifolds, object images, text, single-cell transcriptomes and network logits, a single fixed configuration of COMET ranks first on 21 of the 24 comparisons measured by the three global criteria introduced in Section 3.2, against seven established methods.

Existing methods also expose tunable hyperparameters that can substantially change the embedding. Chosen by eye they carry the analyst's expectations into the embedding, and chosen against a validation measure they carry that measure's criteria. COMET eliminates this source of bias by removing the per-dataset choice. Its parameters are either fixed to defaults designed to work across different datasets or automatically derived from the data. The structure a COMET embedding displays can then be attributed directly to the data rather than to the analyst or to a validation measure.

Several limitations follow from COMET's design and implementation. COMET does not capture sampling density. The locally adaptive kernel in equation 1 normalizes each neighborhood by its own scale, which decouples manifold geometry from how densely a region is sampled, but it also discards density as a signal. Where density is itself the quantity of interest, methods built to preserve it, such as densMAP (Narayan et al., 2021), remain the appropriate choice. COMET also resolves the trustworthiness–continuity tradeoff toward continuity, accepting local intrusions rather than severing genuine neighborhoods. Finally, the current implementation offers no out-of-sample extension. However, this is not intrinsic to the objective: as in parametric UMAP (Sainburg et al., 2021), an encoder trained against the same target probabilities would place new points without recomputation.

Beyond COMET, the measures of Section 3.2 apply to any embedding. Fragmentation and scale distortion are invisible to the rank based statistics this literature reports, yet both appear routinely in the embeddings examined here. Reporting them alongside trustworthiness and continuity would be worthwhile, whatever method produced the embedding.

AI USE STATEMENT

In accordance with ICLR guidelines on Large Language Model (LLM) usage, we disclose that AI tools — including Claude Opus 5, GPT 6 Astra and Gemini 3.1 Pro — were utilized to assist in the research and preparation of this manuscript. Specifically, we used AI in the following capacities:

Research Ideation and Execution: LLMs functioned as an interactive research assistant to help refine experimental protocols, write and optimize codebase structure, and generate data visualization plots.

Retrieval and Discovery: AI tools assisted in surveying academic literature and identifying relevant related work.

Writing and Editing: AI was used to polish and refine the manuscript's prose to improve overall clarity and readability.

In all instances, the authors independently verified the correctness of all AI-generated text, code, and visualizations. The AI tools did not autonomously design the research, select results, or draw scientific conclusions. As per conference policy, the authors assume full and ultimate responsibility for the originality, accuracy, and validity of all contents within this submission.

REPRODUCIBILITY STATEMENT

The method is specified in full in Sections 2.1–2.5, with the derivations of its two data-derived quantities in Appendices C.1 and D.1, and the remaining implementation choices — bridging, initialization and its fallback, and resampling — in Appendices C.2, E.1 and E.2. The single configuration used for every experiment is given in Section 3.1. All five validation measures are defined in Section 3.2 and specified in full in Appendix A, including the landmark estimators used on the atlas dataset and the spanning-tree substitution used above $n = 5000$. The machine and the version of every library used are given in Appendix B.6.

Every reported validation measure is a mean over ten independent runs, with standard errors in Appendix B. Those runs are unseeded, since not every method exposes a random seed and seeding only those that do would make them incomparable. A rerun should therefore reproduce the reported means to within their standard errors, rather than the individual numbers. An anonymized repository containing COMET, the validation code and the datasets accompanies the submission.

REFERENCES

Ehsan Amid and Manfred K. Warmuth. Trimap: Large-scale dimensionality reduction using triplets. *CoRR*, abs/1910.00204, 2019. URL http://arxiv.org/abs/1910.00204.

B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A First Course in Order Statistics*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1992. ISBN 9780898719062. URL https://books.google.se/books?id=dVP-RTea5wcC.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2006.04.006. URL https://www.sciencedirect.com/science/article/pii/S1063520306000546. Special Issue: Diffusion Maps and Wavelets.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numer. Math.*, 1(1):269–271, December 1959. ISSN 0029-599X. doi: 10.1007/BF01386390. URL https://doi.org/10.1007/BF01386390.

Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 577–586, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306324. doi: 10.1145/1963405.1963487. URL https://doi.org/10.1145/1963405.1963487.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. ISSN 08940347, 10886834. URL https://www.jstor.org/stable/jamermathsoci.29.4.983.

Francois Fouss, Alain Pirotte, Jean-michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007. doi: 10.1109/TKDE.2007.46.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Hyeon Jeon, Aeri Cho, Jinhwa Jang, Soohyun Lee, Jake Hyun, Hyung-Kwon Ko, Jaemin Jo, and Jinwook Seo. Zadu: A python library for evaluating the reliability of dimensionality reduction embeddings. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 196–200, 2023. doi: 10.1109/VIS54172.2023.00048.

Dmitry Kobak and George C. Linderman. Umap does not preserve global structure any better than t-sne when using the same initialization. *bioRxiv*, 2019. URL https://api.semanticscholar.org/CorpusID:213746683.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide*. Society for Industrial and Applied Mathematics, 1998. doi: 10.1137/1.9780898719628. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898719628.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, Dec 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0336-3. URL https://doi.org/10.1038/s41587-019-0336-3.

Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, 39(6):765–774, Jun 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00801-7. URL https://doi.org/10.1038/s41587-020-00801-7.

Samer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20). Technical Report CUCS-005-96, Department of Computer Science, Columbia University, February 1996.

Jaroslav Nešetřil, Eva Milková, and Helena Nešetřilová. Otakar borůvka on minimum spanning tree problem translation of both the 1926 papers, comments, history. *Discrete Mathematics*, 233(1):3–36, 2001. ISSN 0012-365X. doi: https://doi.org/10.1016/S0012-365X(00)00224-7. URL https://www.sciencedirect.com/science/article/pii/S0012365X00002247. Czech and Slovak 2.

Jonathan S. Packer, Qin Zhu, Chau Huynh, Priya Sivaramakrishnan, Elicia Preston, Hannah Dueck, Derek Stefanik, Kai Tan, Cole Trapnell, Junhyong Kim, Robert H. Waterston, and John I. Murray. A lineage-resolved molecular atlas of ¡i¿c. elegans¡/i¿ embryogenesis at single-cell resolution. *Science*, 365(6459):eaax1971, 2019. doi: 10.1126/science.aax1971. URL https://www.science.org/doi/abs/10.1126/science.aax1971.

Miloš Radovanovi, Alexandros Nanopoulos, and Mirjana Ivanovi. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010. URL http://jmlr.org/papers/v11/radovanovic10a.html.

Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 10 2021. ISSN 0899-7667. doi: 10.1162/neco_a_01434. URL https://doi.org/10.1162/neco_a_01434.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

The PLOS Computational Biology Staff. Correction: The specious art of single-cell genomics. *PLOS Computational Biology*, 21(11):1–1, 11 2025. doi: 10.1371/journal.pcbi.1013655. URL https://doi.org/10.1371/journal.pcbi.1013655.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2319.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889–899, 2006. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2006.05.014. URL https://www.sciencedirect.com/science/article/pii/S0893608006000724. Advances in Self Organising Maps - WSOM'05.

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(13):451–490, 2010. URL http://jmlr.org/papers/v11/venna10a.html.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z. URL https://doi.org/10.1007/s11222-007-9033-z.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Yingfan Wang, Yiyang Sun, Haiyang Huang, and Cynthia Rudin. Dimension reduction with locally adjusted graphs. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i20.35436. URL https://doi.org/10.1609/aaai.v39i20.35436.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL http://distill.pub/2016/misread-tsne.

Tobias Wängberg, Joanna Tyrcha, and Chun-Biu Li. Shape-aware stochastic neighbor embedding for robust data visualisations. *BMC bioinformatics*, 23(1):477, 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-05028-8.

A Validation Measures

Throughout, $x_1, \ldots, x_n$ is the high-dimensional (HD) data and $y_1, \ldots, y_n$ the low-dimensional (LD) embedding. Every k-nearest-neighbor graph below is built the same way: each point is joined to its k nearest neighbors, each edge is weighted by the Euclidean distance between its endpoints, and an edge is kept if either endpoint lists the other.

Geodesic rank correlation ρ_{Geo}. Shortest paths are computed between all $\binom{n}{2}$ pairs of nodes in both the HD space and the LD space. This yields two vectors of distances, one for each space, that are compared via their Spearman correlation ρ_{Geo}.

Both graphs are bridged before any path is computed (Appendix C.2). The alternative would be to compute the paths first and discard the pairs that come back infinite, and those pairs are not missing at random: they are exactly the pairs the embedding has separated. A method that fragments the data would lose more of them and be scored on the smaller and more forgiving subset that survives, while a method that keeps the manifold in one piece would be scored on everything. The number of surviving pairs would differ from method to method as well, so the correlations would no longer be comparable.

Two steps are carried out differently on the atlas dataset to make it computationally tractable. The first is bridging. COMET's rule (Appendix C.2) takes each component in turn, searches the whole complement for that component's shortest outgoing edge, and adds one such edge per component per round until the graph is connected. Inside COMET this runs on the HD graph, which at $k = 15$ tends to be connected or nearly so, giving few components and few rounds. Here the same rule would run on embeddings that have already broken into as many as 188 pieces, which is the expensive case. The components are therefore joined along a minimum spanning tree of their centroids, computed in a single pass, and each pair of components that tree selects is joined by the exact closest pair of points between them. Both rules add $c - 1$ edges to connect c components. They only differ in which components get paired: COMET's rule pairs them by their true shortest connecting edge, this one by the distance between their centroids. The second optimization is the all-pairs distance matrix, which at $n = 86{,}024$ would hold 7.4×10^9 entries. Shortest paths are instead computed from $L = 1000$ source points instead of from all n. The sources are drawn once with a fixed seed, the same ones are used for the data and for every embedding, and ρ_{Geo} is the Spearman correlation over the $L \times n$ source-to-point distances.

Component count N_{comp}. Recall that N_{comp} counts the components of the embedding's $k^\star$-NN graph, where $k^\star = \max\{5, k_0\}$ and k_0 is the smallest neighborhood size at which the data's own graph is connected. Finding k_0 is cheap, since the number of components never increases with k: a binary search over $k \le 120$ locates it. That ceiling is never reached on these datasets. The largest k_0 among them is COIL-20's 52, which leaves a wide margin.

The floor of 5 is there because k_0 can be very small. Data that is locally one-dimensional connects as soon as each point can reach its immediate neighbors along the curve, and the branching tree does so at $k_0 = 2$. Scoring the embeddings at $k = 2$ would be uninformative, because at that resolution they all come apart, COMET into 59 pieces and the worst method into 167. Raising the floor removes the differences between methods from the other side: at a floor of 15, COMET and PHATE return the same count on every synthetic dataset. Five sits between these two failures. It is fixed before any embedding is scored and applied to every dataset, so it is not tuned to any of them. It only binds where the data connects at fewer than five neighbors, which here means the branching tree and the atlas, whose k_0 is 3. Everywhere else $k^\star = k_0$, as with 20 Newsgroups at 10 and COIL-20 at 52.

Scale distortion D_{scale}. Let $\mathcal{T}$ be the edge set of the data's Euclidean minimum spanning tree, so $|\mathcal{T}| = n - 1$, and for an edge $e = \{i, j\} \in \mathcal{T}$ let $a_e = \|x_i - x_j\|$ and $b_e = \|y_i - y_j\|$ be its length in the data and in the embedding.[5] Normalizing each side gives the edge shares

$$P_e = \frac{a_e}{\sum_{e' \in \mathcal{T}} a_{e'}}, \qquad Q_e = \frac{b_e}{\sum_{e' \in \mathcal{T}} b_{e'}}, \qquad e \in \mathcal{T},$$

so that P_e is the fraction of the data's total tree length carried by edge e, and Q_e the fraction the embedding gives the same pair of points. The reported value is

$$D_{\mathrm{scale}} \;=\; D_{\mathrm{KL}}(P \,\|\, Q) \;=\; \sum_{e \in \mathcal{T}} P_e \log \frac{P_e}{Q_e}. \tag{7}$$

[5] An embedding can place two distinct points close enough that b_e underflows to zero in double precision, which PHATE does on 231 of the atlas's $86{,}023$ edges, and a single $b_e = 0$ would make equation 7 infinite. Lengths are therefore floored at $\varepsilon_{\mathrm{mach}} \max_{e'} b_{e'}$ before normalizing, where $\varepsilon_{\mathrm{mach}} \approx 2.2 \times 10^{-16}$ is the machine epsilon.

The comparison is made edge by edge rather than between summaries of the two distributions. A summary is too coarse for what this measure has to catch: an embedding that shortens one group of edges and lengthens another by matching amounts leaves Q spread as evenly as P, so any comparison of spread alone would score it near zero although the tree has been rearranged. In equation 7 each edge's share is weighed against its own counterpart, and the result is zero only when no edge has changed its share of the total tree length.

Since the KL-divergence is asymmetric the order of the arguments P and Q matter here and it follows the usual convention. $D_{\mathrm{KL}}(P \,\|\, Q)$ measures the information lost when Q is used to approximate P, so P is the reference and Q the approximation to it. The data's edge shares are the reference here and the embedding's are the approximation, which fixes the order in equation 7.

Building $\mathcal{T}$ only requires distances in the HD space. Below $n = 5000$ the tree is exact, built from the full pairwise distance matrix. That matrix is 59 GB at the atlas's $n = 86{,}024$, so above the threshold $\mathcal{T}$ is the minimum spanning tree of a 60-nearest-neighbor graph instead. Restricting the candidate edges in this way changes the tree only if the exact tree uses an edge that the 60-nearest-neighbor graph leaves out. On the synthetic datasets, where both trees can be built, that never happens. The graph is also tested for connectivity before the tree is built, since a disconnected graph yields a forest of fewer than $n - 1$ edges and would change both normalizations.

Trustworthiness. Let $\mathcal{K}_i$ and $\widehat{\mathcal{K}}_i$ be the k nearest neighbors of i in the data and in the embedding, respectively, and let $r(i, j)$ be the rank of j when the other $n - 1$ points are ordered by their ambient distance from x_i, so that $r = 1$ is the nearest. The intrusions at i are $\mathcal{I}_i = \widehat{\mathcal{K}}_i \setminus \mathcal{K}_i$, the points the embedding has brought into the neighborhood of i that the data leaves out. Each is charged the rank it holds in the data, minus k, so that a point just outside the true neighborhood costs almost nothing and a point from across the manifold costs a great deal (Venna & Kaski, 2006):

$$\mathrm{Trust}(k) \;=\; 1 - \alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{I}_i} \big(r(i,j) - k\big), \qquad \alpha = \frac{2}{nk(2n - 3k - 1)}. \tag{8}$$

The constant α is the reciprocal of the largest value the double sum can take, which places $\mathrm{Trust}(k)$ in $[0, 1]$ with 1 meaning no intrusions at all.

Continuity. The extrusions at i are $\mathcal{E}_i = \mathcal{K}_i \setminus \widehat{\mathcal{K}}_i$, the points the data places in the neighborhood of i that the embedding leaves out. Each is charged its rank $\hat{r}(i, j)$ in the embedding, again minus k, which is equation 8 with the roles of the two spaces exchanged:

$$\mathrm{Cont}(k) \;=\; 1 - \alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{E}_i} \big(\hat{r}(i,j) - k\big). \tag{9}$$

Both are computed at $k = 15$ with the `zadu` implementation (Jeon et al., 2023).

Ranking all n points from all n query points is $O(n^2)$, which is out of reach on the atlas. The outer sums in equation 8 and equation 9 run there over the same $L = 1000$ landmarks used for ρ_{Geo}, and α becomes $2/\big(Lk(2n - 3k - 1)\big)$. Each source still ranks every one of the n points, so the ranks entering the sums are exact and only the average over query points is sampled.

B Complete Results

B.1 Trustworthiness and Continuity

	Flat disc		Swiss roll		Hier. Gauss.		Tree	
Method	Trust	Cont	Trust	Cont	Trust	Cont	Trust	Cont
COMET	**0.9998**	**0.9998**	**0.9997**	**0.9998**	0.9748	**0.9918**	0.9860	0.9892
t-SNE	0.9993	0.9983	**0.9997**	0.9942	**0.9937**	0.9889	**0.9914**	**0.9913**
UMAP	0.9984	0.9964	0.9995	0.9986	0.9876	0.9896	0.9885	0.9872
PHATE	0.9677	0.9911	0.9493	0.9962	0.9629	0.9784	0.9697	0.9864
TriMAP	0.9987	0.9988	0.9844	0.9984	0.9850	0.9900	0.9889	0.9908
PaCMAP	0.9989	0.9987	0.9975	0.9939	0.9784	0.9892	0.9883	0.9889
SASNE	0.9968	0.9942	0.9762	0.9918	0.9611	0.9742	0.9843	0.9882
LocalMAP	0.9980	0.9926	0.9992	0.9914	0.9900	0.9826	0.9887	0.9874

Table 3: **Trustworthiness and continuity, averaged over 10 independent runs.** Venna–Kaski Trustworthiness and Continuity at $k = 15$, both higher-is-better, on the four synthetic datasets. Bold values mark the best mean per column and any method within one standard error of the difference from it.

	COIL-20		20NG		*C. elegans*		CIFAR-10	
Method	Trust	Cont	Trust	Cont	Trust	Cont	Trust	Cont
COMET	0.9773	**0.9958**	0.8291	**0.9563**	0.9699	**0.9929**	0.9782	**0.9917**
t-SNE	**0.9931**	0.9915	**0.9817**	0.9427	**0.9969**	0.9896	**0.9928**	0.9895
UMAP	0.9915	0.9934	0.9590	0.9456	0.9889	0.9904	0.9841	0.9907
PHATE	0.8982	0.9749	0.7408	0.9085	0.9358	0.9858	0.9495	0.9769
TriMAP	0.9841	0.9876	0.8569	0.9305	0.9743	0.9908	0.9794	0.9905
PaCMAP	0.9820	0.9881	0.8579	0.9251	0.9811	0.9885	0.9796	0.9877
SASNE	0.9601	0.9869	0.7279	0.8789	–	–	0.9754	0.9870
LocalMAP	0.9920	0.9852	0.9459	0.9174	0.9913	0.9863	0.9862	0.9866

Table 4: **Trustworthiness and continuity on the real datasets, averaged over 10 independent runs.**

B.2 Standard Errors

	Flat disc			Swiss roll			Hier. Gaussians			Tree		
Method	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}
COMET	0.0001	0.00	0.0005	0.0001	0.00	0.0009	0.0020	0.13	0.0014	0.0004	0.22	0.0020
t-SNE	0.0000	0.00	0.0000	0.0000	0.00	0.0000	0.0000	0.00	0.0000	0.0000	0.00	0.0000
UMAP	0.0165	0.86	0.0029	0.0023	1.17	0.0050	0.0292	0.00	0.0040	0.0253	0.52	0.0075
PHATE	0.0005	0.29	0.0016	0.0025	1.22	0.0065	0.0362	0.00	0.0021	0.0001	0.62	0.0025
TriMAP	0.0001	0.00	0.0016	0.0015	0.40	0.0012	0.0025	0.26	0.0029	0.0001	0.48	0.0049
PaCMAP	0.0004	0.18	0.0022	0.0080	0.23	0.0076	0.0045	0.00	0.0057	0.0019	0.93	0.0088
SASNE	0.0000	0.00	0.0000	0.0000	0.00	0.0000	0.0000	0.00	0.0000	0.0000	0.00	0.0000
LocalMAP	0.0140	0.60	0.0035	0.0039	0.37	0.0099	0.0394	0.48	0.0180	0.0129	0.79	0.0082

Table 5: **Standard errors of Table 1.**

	COIL-20			20NG			*C. elegans*			CIFAR-10		
Method	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}	ρ_{Geo}	N_{comp}	D_{scale}
COMET	0.0065	0.20	0.0033	0.0002	0.00	0.0011	0.0023	0.43	0.0010	0.0005	0.00	0.0012
t-SNE	0.0000	0.00	0.0000	0.0000	0.00	0.0000	0.0000	0.00	0.0000	0.0000	0.00	0.0000
UMAP	0.0121	0.20	0.0111	0.0037	0.43	0.0065	0.0173	3.48	0.0099	0.0255	0.64	0.0040
PHATE	0.0058	0.20	0.0114	0.0038	0.31	0.0107	0.0068	10.83	0.0132	0.0110	0.48	0.0110
TriMAP	0.0070	0.13	0.0058	0.0006	0.23	0.0069	0.0008	0.56	0.0012	0.0008	0.31	0.0025
PaCMAP	0.0122	0.15	0.0134	0.0043	0.46	0.0067	0.0156	1.96	0.0061	0.0122	0.34	0.0067
SASNE	0.0000	0.00	0.0000	0.0000	0.00	0.0000	–	–	–	0.0000	0.00	0.0000
LocalMAP	0.0132	0.47	0.0097	0.0118	0.89	0.0068	0.0266	2.36	0.0068	0.0212	0.41	0.0035

Table 6: **Standard errors of Table 2.**

	Flat disc		Swiss roll		Hier. Gauss.		Tree	
Method	Trust	Cont	Trust	Cont	Trust	Cont	Trust	Cont
COMET	0.00000	0.00000	0.00002	0.00001	0.00016	0.00004	0.00005	0.00005
t-SNE	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
UMAP	0.00001	0.00005	0.00000	0.00007	0.00021	0.00004	0.00003	0.00009
PHATE	0.00010	0.00001	0.00134	0.00006	0.00007	0.00001	0.00013	0.00002
TriMAP	0.00001	0.00001	0.00025	0.00002	0.00022	0.00004	0.00003	0.00002
PaCMAP	0.00001	0.00001	0.00038	0.00022	0.00025	0.00007	0.00003	0.00010
SASNE	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
LocalMAP	0.00003	0.00008	0.00007	0.00016	0.00004	0.00030	0.00006	0.00024

Table 7: **Standard errors of Table 3.**

	COIL-20		20NG		*C. elegans*		CIFAR-10	
Method	Trust	Cont	Trust	Cont	Trust	Cont	Trust	Cont
COMET	0.00072	0.00004	0.00032	0.00002	0.00039	0.00006	0.00002	0.00002
t-SNE	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
UMAP	0.00041	0.00016	0.00031	0.00038	0.00020	0.00019	0.00006	0.00003
PHATE	0.00148	0.00066	0.00299	0.00256	0.00291	0.00034	0.00250	0.00098
TriMAP	0.00162	0.00017	0.00036	0.00052	0.00026	0.00004	0.00006	0.00004
PaCMAP	0.00085	0.00023	0.00094	0.00062	0.00021	0.00028	0.00011	0.00007
SASNE	0.00000	0.00000	0.00000	0.00000	–	–	0.00000	0.00000
LocalMAP	0.00011	0.00046	0.00072	0.00043	0.00013	0.00024	0.00007	0.00004

Table 8: **Standard errors of Table 4.**

B.3 Ablation of the commute-time target probabilities

Commute times enter COMET in two places: the target probabilities of the sampled non-neighbors, which are ordered by commute-time rank, and the spectral initialization, which is the commute-time embedding itself. To check that the target probabilities contribute something the initialization alone does not, we ablate the two independently on the hierarchical Gaussians. *Spectral init* is standard COMET. *PCA init* keeps the calibrated target probablities but starts from the first two principal components, the same start COMET falls back under the conditions detailed in Appendix E.1. *Uniform repulsion* sets the target probability of every non-neighbor to zero, $p_t = 0$ for $t > k$, so all non-neighbors are repelled equally. It is combined with each initialization, giving four combinations. The embeddings for each combination are shown in Figure 4. Each configuration is run 10 times, unseeded, and scored with the global measures of the main text (Table 9).

Replacing the commute time ordered target probabilities with 0 raises D_{scale} by two thirds and causes over-separation of the clusters giving five or six components, compared to around 2 with the targets present. Meanwhile Trustworthiness increases slightly suggesting the commute-ordering improves global structure preservation at a small cost in local structure preservation. The initialization makes barely any difference on this dataset once the targets are present. The commute ordering of the targets,

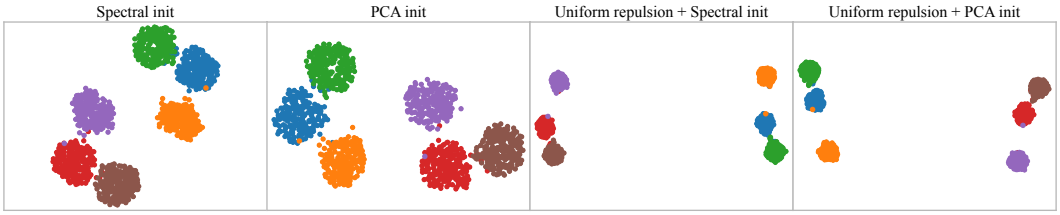


Figure 4: **Commute-ordering ablation on the hierarchical Gaussians.** One embedding for each combination. With COMETs target probabilities (left two panels) the hierarchical organization is recovered from either initialization. With every non-neighbor's target probability set to 0 (right two panels) the clusters stay separated but their relative arrangement is lost.

not the initialization, is what preserves the hierarchical arrangement of the clusters and their relative distances.

Configuration	ρ_{Geo}	N_{comp}	D_{scale}	Trust	Cont
Spectral init	0.966 $\pm$0.007	2.1 $\pm$0.1	0.163 $\pm$0.002	0.9749 $\pm$0.0001	0.9918 $\pm$0.0000
PCA init	0.969 $\pm$0.007	2.1 $\pm$0.1	0.159 $\pm$0.002	0.9749 $\pm$0.0002	0.9917 $\pm$0.0001
Uniform repulsion + Spectral init	0.948 $\pm$0.017	5.1 $\pm$0.2	0.269 $\pm$0.003	0.9771 $\pm$0.0001	0.9916 $\pm$0.0001
Uniform repulsion + PCA init	0.964 $\pm$0.003	5.2 $\pm$0.2	0.263 $\pm$0.004	0.9773 $\pm$0.0001	0.9915 $\pm$0.0000

Table 9: **Validation measures for the Commute-ordering ablation.** The four configurations of Figure 4 on the hierarchical Gaussians; mean over 10 independent runs, $\pm$ standard error.

B.4 Sensitivity to the non-neighbor sample size s

Table 10 varies s from 5 to 100, holds everything else at its default, and reports two things: the five validation measures on the Swiss roll and the hierarchical Gaussians, each the mean of ten independent runs, and the wall-clock time COMET needs to embed the full *C. elegans* atlas, the mean of three runs. The measures are computed by the same code as the main tables, so the $s = 15$ row differs from Table 1 only by the run-to-run spread.

Increasing s improves some measures and degrades others. On the Swiss roll, ρ_{Geo} rises and D_{scale} falls until $s = 50$, after which both reverse. On the hierarchical Gaussians, ρ_{Geo} and D_{scale} are at their best near $s = 25$, while the embedding fragments steadily as s grows, from 2.0 components at $s = 5$ to 4.2 at $s = 100$. Trustworthiness improves slowly with s on the Gaussians, and continuity is flat. No value of s is best on every measure.

Runtime is monotone in s, and it rises steeply. Embedding the atlas takes 22.0 s at the default $s = 15$, 96.1 s at $s = 50$ and 336.3 s at $s = 100$, so raising s above 15 multiplies the time by between four and fifteen. That growth is close to linear in s up to about $s = 25$ and steeper above it, consistent with the eigensolve, whose cost grows faster than linearly in $l = s$. Note that this is the scaling in s; for a fixed small s the scaling in n is linear as discusses in Section 2.6 . Lowering s below $s = 15$ returns far less in terms of performance gains and comes at a significant cost in embedding quality: $s = 5$ takes 15.5 s and doubles the scale distortion on the Swiss roll.

The default $s = 15$ rests on that imbalance. Between $s = 15$ and $s = 50$ the validation measures move little, while runtime grows more than fivefold and climbs sharply above that range. The value $s = 15$ is not the best on any single measure, but it sits inside the flat region of most measures and below the point at which runtime turns steeply upward.

	Swiss roll					Hier. Gaussians					*C. elegans*	
s	ρ_{Geo}	N_{comp}	D_{scale}	Trust	Cont	ρ_{Geo}	N_{comp}	D_{scale}	Trust	Cont	Fit (s)	Rel.
5	0.969	1.0	0.183	0.9981	0.9991	0.967	2.0	0.185	0.9723	0.9900	15.5	0.71×
10	0.974	1.0	0.126	0.9994	0.9997	0.973	2.0	0.169	0.9741	0.9912	17.1	0.78×
15	0.980	1.0	0.090	0.9998	0.9998	0.971	2.2	0.161	0.9748	0.9918	22.0	1.00×
25	0.986	1.0	0.063	0.9999	0.9999	0.979	2.5	0.158	0.9757	0.9921	33.0	1.50×
50	0.989	1.0	0.047	0.9999	0.9999	0.976	3.6	0.163	0.9783	0.9924	96.1	4.38×
100	0.950	1.0	0.068	0.9996	0.9996	0.975	4.2	0.176	0.9831	0.9924	336.3	15.31×

Table 10: **Sensitivity to the non-neighbor sample size s.** Validation measures are the mean over 10 independent runs at each s, with every other setting at its default; $l = s$ throughout, so the spectral dimension varies with s. The default $s = 15$ is in bold. The final columns are the wall-clock time to embed the full *C. elegans* atlas ($n = 86{,}024$), in seconds and relative to $s = 15$.

B.5 Correlation between ambient and shortest-path distances

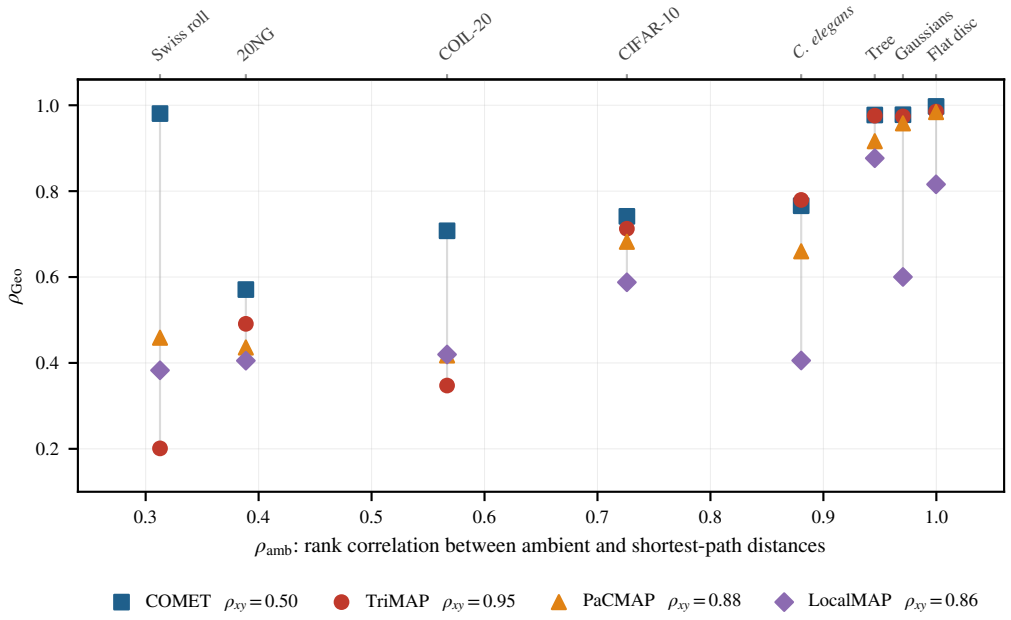


Figure 5: **Geodesic rank correlation tracks ambient–shortest-path agreement for the methods that rely on ambient distances beyond the local scale, and not for COMET.** Each column is one dataset and each marker one method applied to it, the four joined by a faint line, with ρ_{Geo} from Tables 1 and 2. The horizontal axis is ρ_{amb}, the Spearman correlation between ambient and shortest-path distances over all pairs, using a subsample of $4{,}000$ points when $n > 4000$. Shortest paths are computed on a 15-NN graph, and ρ_{amb} is a property of the dataset alone, computed before any embedding exists. PaCMAP, TriMAP and LocalMAP sample or weight their pairs by ambient distance, and for all three ρ_{Geo} follows ρ_{amb} at rank correlations above 0.85. On the three datasets where the two orderings agree least — the Swiss roll, COIL-20 and 20 Newsgroups — they reach ρ_{Geo} between 0.35 and 0.44 on average, against COMET's 0.75. Where the orderings nearly coincide COMET and TriMAP differ by at most 0.014.

B.6 Runtime Analysis

Hardware and software. Every runtime in this paper was measured on one machine: a MacBook Air with an Apple M4 processor and 32 GB of unified memory. Methods were run sequentially in a single process on CPU, with no other significant load, and each library's own threading defaults were left in place, so the comparison reflects the implementations as distributed rather than a controlled single-threaded measurement. Versions were Python `3.12.7`, NumPy `2.4.6`, SciPy `1.18.0`, scikit-learn `1.9.0`, `umap-learn` `0.5.12`, `pacmap` `0.9.1`, `trimap` `1.1.5`[6], `phate` `2.0.0` and `zadu` `0.4.2`. Note that we used the t-SNE implementation in scikit-learn and the implementations of both PacMAP and LocalMAP from the `pacmap` library. For SASNE the official MATLAB implementation was used.

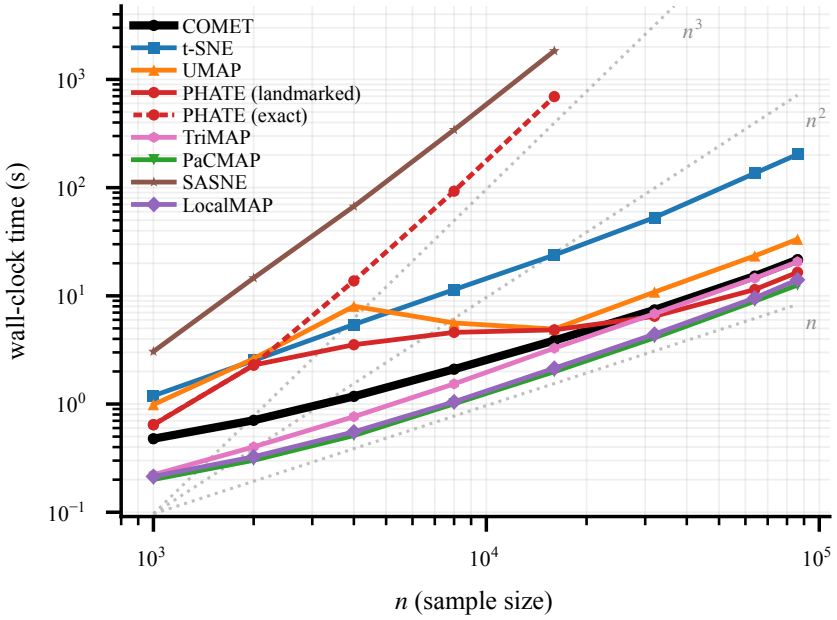


Figure 6: **Wall-clock scaling on nested subsamples of the *C. elegans* atlas.** Time to return a 2D embedding at $n = 1{,}000$, $2{,}000$, $4{,}000$, $8{,}000$, $16{,}000$, $32{,}000$, $64{,}000$ and the full $86{,}024$. Both axes are logarithmic and the dotted guides are gradients n, n^2 and n^3, so slope rather than height is the comparison. COMET fits the lowest slope of any method above $n = 8{,}000$, 0.98 against 1.07–1.09 for PaCMAP, LocalMAP and TriMAP, and embeds the full atlas in 21.6 s. Those three are faster than COMET at small n, but since COMET scales better here the gap narrows as n grows, from $2.4\times$ PaCMAP at $n = 1{,}000$ to $1.7\times$ at the full atlas. SASNE and exact PHATE stop at $n = 16{,}000$, both out of memory. PHATE is drawn twice because its default setting landmarks at $n_{\text{landmark}} = 2000$: that curve flattens because the expensive stage is fixed at $O(2000^2)$ whatever n is, not because the algorithm scales sublinearly. UMAP's curve is not monotone because the default Python implementation switches from exact to approximate neighbor search at $n = 4{,}096$ and reduces its epoch count from 500 to 200 above $n = 10^4$.

C Graph Construction

C.1 Maximum entropy kernel parameters

Fix a node i and let Δ_i denote the dissimilarity from i to a neighbor $v \in \mathcal{N}_i$, treated as a continuous random variable with finite mean. Writing the kernel of Section 2.1 as a function of the dissimilarity,

$$f_{a,\tau}(\delta) \;=\; \exp\!\left[-\frac{\delta - a}{\tau}\right], \qquad \tau > 0,$$

[6]With `annoy` rebuilt from source; the PyPI wheel returns only the query point from every neighbor search on Apple silicon.

so that $f_{a,\tau}(\delta_{ij}) = f_{i\to j}(a,\tau)$, set $W_{a,\tau} = f_{a,\tau}(\Delta_i)$. The weights are required to lie in $(0,1]$, which holds exactly when $a \le \inf \operatorname{supp}(\Delta_i)$.

By the change-of-variables identity for differential entropy (Shannon, 1948, §20), which applies since $f_{a,\tau}$ is strictly monotone,

$$h(W_{a,\tau}) \;=\; h(\Delta_i) + \mathbb{E}\big[\log |f'_{a,\tau}(\Delta_i)|\big]\,, \qquad f'_{a,\tau}(\delta) = -\frac{1}{\tau}\exp\left[-\frac{\delta - a}{\tau}\right],$$

so that

$$h(W_{a,\tau}) \;=\; h(\Delta_i) - \log\tau - \frac{\mathbb{E}[\Delta_i] - a}{\tau}. \tag{10}$$

The shift and the scale are obtained in turn. Entropy is strictly increasing in the shift, $\partial h/\partial a = 1/\tau > 0$, so the constraint binds and $a = \inf \operatorname{supp}(\Delta_i)$: the subtraction in equation 1 is selected by the requirement that the weights are in $(0,1]$, rather than imposed. Differentiating equation 10 with respect to the scale τ yields,

$$\frac{d}{d\tau}h(W_\tau) = -\frac{1}{\tau} + \frac{\mathbb{E}[\Delta_i] - a}{\tau^2}, \qquad \frac{d^2}{d\tau^2}h(W_\tau) = \frac{1}{\tau^2} - \frac{2(\mathbb{E}[\Delta_i] - a)}{\tau^3}.$$

The first derivative vanishes only at $\tau^\star = \mathbb{E}[\Delta_i] - a$, where the second evaluates to $-1/(\mathbb{E}[\Delta_i] - a^2) < 0$, so the maximizer is unique,

$$\tau^\star \;=\; \mathbb{E}[\Delta_i] - \inf \operatorname{supp}(\Delta_i),$$

whenever $\mathbb{E}[\Delta_i] > \inf \operatorname{supp}(\Delta_i)$.

Both population quantities are replaced by their sample counterparts over the neighbor set $\mathcal{N}_i$,

$$\delta_{\min}(i) = \min_{v\in\mathcal{N}_i} \delta_{iv}, \qquad \overline{\delta}_i = \frac{1}{|\mathcal{N}_i|}\sum_{v\in\mathcal{N}_i}\delta_{iv}, \qquad \widehat{\tau} = \overline{\delta}_i - \delta_{\min}(i),$$

which recovers equation 1. Since $\mathcal{N}_i \supseteq \mathcal{K}_i$, the set is not of fixed size across nodes, and both estimates are taken over whichever neighbors the symmetrized graph assigns to i. The existence condition $\mathbb{E}[\Delta_i] > \inf \operatorname{supp}(\Delta_i)$ becomes $\overline{\delta}_i > \delta_{\min}(i)$, which holds as long as every neighbor of i is not equidistant from it, in which case all weights are set to 1.

C.2 Bridging a disconnected graph

Whether $\bar{G}$ is connected at a given k is a property of the data. If it is not, L_{sym} has one zero eigenvalue per component (von Luxburg, 2007), so $\lambda_1 = 0$ and commute times between components are undefined. We therefore ensure $\bar{G}$ is connected by adding bridging edges between components whenever necessary.

The procedure used is Borůvka's algorithm (Nešetřil et al., 2001) run on the graph whose nodes are the connected components of $\bar{G}$ and whose edge weights are the shortest distances between them. Components are found by a connected-components pass over the symmetrized graph. For each component C, every point of C is queried for its exact nearest neighbor in the complement $X \setminus C$ (a KD-tree for $d \le 15$, brute force above; the query is exact in either case, so the selected edge is the true shortest edge leaving C). The shortest of these edges is C's bridge. All bridges are added simultaneously—two components may select the same edge, in which case it is added once—the components are recomputed, and the round repeats until one component remains. Every component merges with at least one other per round, so at most $\lceil \log_2 c \rceil$ rounds are needed for c components. The added edges form a minimum spanning tree over the components: $c - 1$ edges, each the shortest available.[7]

Bridging edges carry their measured distance and pass through equation 1 like any other edge. A bridge joins two points that are not among each other's k nearest neighbors, otherwise they would already be adjacent, so it is at least as long as the kth-nearest-neighbor distance of both endpoints, and its weight is bounded by the weights the endpoints assign their own kth nearest neighbors. Connectivity is restored without adding a strong similarity anywhere. Bridging edges are excluded from the sets $\mathcal{K}_i$: they make two nodes adjacent without making either one of the other's k nearest.

[7] As in any Borůvka construction, this assumes distinct edge weights; ties in floating-point distances are not a practical concern.

D THE COMMUTE-TIME EMBEDDING

D.1 REGULARIZATION OF THE COMMUTE-TIME EMBEDDING

The regularization term ϵ conditions the spectrum that the commute-time embedding inverts. L_{sym} is symmetric positive semidefinite, so its singular values are its eigenvalues and its condition number is

$$\kappa(L_{sym}) = \frac{\lambda_{\max}(L_{sym})}{\lambda_{\min}(L_{sym})}. \tag{11}$$

The trivial eigenpair is discarded, so $\lambda_{\min}$ is the smallest non-trivial eigenvalue λ_1. Truncating the spectrum leaves $\lambda_{\max}$ unavailable, so it is approximated by 2, the largest value an eigenvalue of L_{sym} can take (von Luxburg, 2007),

$$\kappa(L_{sym}) \approx \frac{2}{\lambda_1}. \tag{12}$$

The approximation overestimates κ, since $\lambda_{\max} \leq 2$, so a threshold met by the estimate is met by the true condition number as well.

Regularization shifts the inverted spectrum to $\lambda_i + \epsilon$. Setting the estimate equal to the threshold κ_t, which is 5000 by default,

$$\kappa_t = \frac{2 + \epsilon}{\lambda_1 + \epsilon}, \tag{13}$$

and solving for ϵ gives

$$\epsilon = \max\left(0,\ \frac{2 - \kappa_t \lambda_1}{\kappa_t - 1}\right), \tag{14}$$

which is zero whenever $\lambda_1 \geq 2/\kappa_t$ in which case no regularization is needed.

E OPTIMIZATION

E.1 INITIALIZATION AND ITS FALLBACK

The embedding is initialized from the first m columns of $\widetilde{C}_l$. Three conditions replace this with the first m principal components of X.

A bridged graph is the first. Bridging connects the components through a few long edges, so the smallest eigenpairs of L_{sym} describe that near-partition. This causes the coordinates of $\widetilde{C}_l$ to create a large separation between the data points in these components, making it a poor starting point for gradient descent, which has to spend many epochs to bring the components back together.

The large separation between components from bridging can happen even when bridging was not used if the connections between different regions of the data are weak. To address these cases COMET also switches to PCA initialization if any coordinate of the spectral initialization is stretched by a small number of points. Writing $(\widetilde{C}_l)_{\cdot,j}$ for the j-th column of $\widetilde{C}_l$, the initialization is replaced whenever

$$\max_{1 \leq j \leq m} \frac{\max_i (\widetilde{C}_l)_{ij} - \min_i (\widetilde{C}_l)_{ij}}{\mathrm{sd}\big((\widetilde{C}_l)_{\cdot,j}\big)} \ > \ 7,$$

where sd is the standard deviation over the n points. The ratio is invariant to the scale of $\widetilde{C}_l$, so it does not depend on the normalization applied to it. A coordinate whose values occupy its range takes a small value, $\sqrt{12} \approx 3.5$ for a uniform spread and $2\sqrt{2} \approx 2.8$ for a sinusoidal one, while a coordinate whose extent is set by a handful of isolated points takes a much larger one. The threshold of 7 is roughly twice the value a well-spread coordinate takes.

Too few columns is the last condition where PCA initialization is used. C_l supplies l coordinates, so an embedding of dimension $m > l$ cannot be read from it.

The fallback changes only where gradient descent starts. Negative samples are ranked in C_l throughout, so the commute-time geometry drives the objective either way.

E.2 RESAMPLING THE NON-NEIGHBORS

The s non-neighbors sampled for each anchor cover only a small fraction of the $n-k-1$ available points. By default COMET runs 2000 epochs of asynchronous stochastic gradient descent and redraws the non-neighbors every 10 epochs to broaden coverage and reduce dependence on a particular sampled interaction set. Within each draw, points are ordered by distance in $\widetilde{C}_l$, and the t-th point receives the target probability associated with the expected population rank $r_t = k+t(n-k)/(s+1)$. A particular point can occupy different positions across draws and therefore receive different targets. Resampling thus optimizes a stochastic rank-based objective rather than recovering each pair's exact population-rank target. Excluding $\mathcal{K}_i$ and the anchor keeps the sampling pool at $n-k-1$ for every node, making the same order-statistic formula applicable throughout.