Automating t-SNE parameterization with prototype-based learning of manifold connectivity

Josh Taylor [a,*], Erzsébet Merényi [b]

[a] *The University of Texas at Austin, Austin, TX, USA*
[b] *Rice University, Houston, TX, USA*

ABSTRACT

We harness topological information about a data manifold revealed through neural prototype-based learning to automate t-SNE parameterization. This information is contained in the CONN (CONNectivity) similarity of neural prototypes, which grades the strength (weakness) of topological connectivity at various points within a data manifold. CONN suggests a data-driven specification of localized versions (varying across the manifold) of t-SNE's perplexity parameter which, in turn, defines the high-dimensional similarities P that t-SNE attempts to preserve. We further imbue P with CONN's graded similarity to reduce mismatch between the topology of the manifold and its embedded representation. Experiments show these improvements, collectively called **CONNt-SNE**, are capable of producing meaningful and trustworthy low-dimensional embeddings without the need to heuristically optimize over (i.e., grid search) t-SNE's perplexity space. Data-driven t-SNE parameterization improves our confidence that any structure appearing in the embeddings is valid and not merely an artifact of spurious parameterization.

© 2022 Published by Elsevier B.V.

1. Introduction

As a sub-field of unsupervised machine learning, dimensionality reduction seeks to represent points in high-dimensional data clouds by points in lower-dimensional spaces as faithfully as possible. Linear projections from high-d to low-d, such as Principal Component Analysis, are commonly used for this task, but cannot faithfully represent non-linear high-d structures. As such, many nonlinear dimensionality reduction (NDR) techniques [1] have been proposed to produce more accurate low-d representations of complex high-d manifolds. However, like most of machine learning, poor parameterization can greatly affect the quality of NDR techniques.

This work presents a data-driven framework we call **CONNt-SNE** for automatically parameterizing a widely used NDR algorithm known as t-SNE (**t**-distributed **S**tochastic **N**eighbor **E**mbedding, [2]). The framework is intended for, and based upon knowledge obtained from, prototype-based manifold learning. In Section 2 we briefly review the t-SNE algorithm and concepts from prototype-based learning which form the basis for CONNt-SNE's

methodology, outlined in Section 3. In Section 4 we review existing measures used to assess the quality of NDR techniques, discuss their strengths and weaknesses, and contribute an additional measure of topology preservation known as TNE (**T**opological **N**eighborhood **E**xpansion) to this canon. Section 5 details the experimental design underpinning the results of Section 6, which show CONNt-SNE meets or exceeds the performance of regular t-SNE, even when its performance is optimized over a grid of possible parameter values. We stress here that such optimization of t-SNE's parameters is only possible in controlled experimental settings where additional knowledge of high-d data structure is available (e.g., through a priori known cluster structure or manifold structure). For most use-cases in unsupervised learning such external information is unavailable, which motivated development of CONNt-SNE as a way to provide trustworthy, automated t-SNE parameterizations.

2. Background

t-SNE [2] has attracted wide attention both within and outside the machine learning community as a tool for producing low-dimensional non-linear embeddings $T = \{t_s \in \mathbb{R}^{d'}\}_{s=1}^{N}$ of high-dimensional point clouds $X = \{x_s \in \mathbb{R}^d\}_{s=1}^{N}$, where $d' \ll d$, for

---

* Corresponding author.
E-mail addresses: joshtaylor@utexas.edu (J. Taylor), erzsebet@rice.edu (E. Merényi).

exploratory (visual) data analysis. Typically $d' \in \{2, 3\}$. The appetite for such analysis across disciplines is strong, but many questions have been raised about what, exactly, can (should) be inferred from a t-SNE embedding. t-SNE's introduction subtly stresses its distinction as a technique for visualization (vs. feature engineering), yet its embeddings are often clustered either informally (via visual assessment) or formally (applying a clustering algorithm to T). Some [3] have noticed relative deficiencies in t-SNE's ability to faithfully indicate separation in complex manifolds. [4] offers a list of various misinterpretations that can be made from a t-SNE embedding due to its unfaithful representation of cluster sizes, shapes, densities, compactness and separability. Most of these issues arise because t-SNE is designed to preserve conditional probabilities between points instead of distance, and we believe they are not severe impediments to successful cluster discovery from low-d representations. Indeed, over the last three decades the lattice representations of data learned by Self-Organizing Maps [5] have produced many successful clusterings without explicit preservation of, e.g. distance, between the high- and low-d spaces. However, [4] does raise one issue we feel fundamentally impacts the fidelity of a t-SNE representation: that of selecting its main perplexity parameter, which we abbreviate px. px indirectly controls the number of Euclidean neighbor similarities that t-SNE attempts to preserve, which is an unknown number that varies across, and likely within, datasets. An example taken from [4] of various t-SNE embeddings which can arise from different px specifications is given in Fig. 1. Here, the "high-d" data (left-most panel) is very simple—two dimensional with two well-defined clusters—yet inspection of the embeddings resulting from some perplexity values (2, 5, 100) would yield a different conclusion. [2] suggests that t-SNE is relatively insensitive to px but in practice an optimal perplexity is obviously data-dependent and should be data-driven. CONNt-SNE provides a mechanism for such a scheme, using information freely available from prototype-based learning, and commonly invoked during prototype-based clustering.

2.1. The t-SNE algorithm

The t-SNE algorithm begins by defining Gaussian similarities between two points in $\mathbb{R}^d$ as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \quad p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_k - x_i\|^2/2\sigma_i^2)} \tag{1}$$

where $p_{\cdot|i}$ is the conditional distribution of all other x_j given x_i and, by convention, $p_{i|i} = 0$. We let $P = \{p_{ij}\}$ be the $N \times N$ matrix of such (symmetrized) similarities and denote its i-th row by P_i. Each Gaussian bandwidth σ_i is controlled by the (global) perplexity parameter px, found through iterative search such that following relationship holds:

$$\sigma_i : px = 2^{H(P_i)}, \quad H(P_i) = -\sum_j p_{j|i} \log_2(p_{j|i}). \tag{2}$$

Pointwise similarities q_{ij} in $\mathbb{R}^{d'}$ are derived from the pdf of Student's t-distribution with one degree of freedom,

$$q_{ij} = \frac{(1 + \|t_i - t_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|t_k - t_l\|^2)^{-1}}, \tag{3}$$

where again we let $Q = \{q_{ij}\}$. Embedded coordinates t_i are determined through minimization of the Kullback–Leibler divergence as cost,

$$C = KL(P\|Q) = \sum_{ij} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right). \tag{4}$$

2.2. CONN similarity

CONNt-SNE provides a framework for embedding the *prototypes* $W = \{w_i \in \mathbb{R}^d\}_{i=1}^M, M \ll N$, of a vector quantizer (VQ) trained on data X. While the prototypes of any VQ would be suitable for this purpose we prefer neural variants such as the SOM and Neural Gas (NG, [6]) as the iterative stages of competition and cooperation during training result in better prototype placement in the data cloud than, e.g., k-means [7]. Previous work [8] utilized t-SNE as a means to visualize Neural Gas prototypes but, contrary to this work, did not explore any ways by which t-SNE could be influenced by the VQ. To achieve the latter we appeal to the CONN similarity [9] between trained prototypes w_i and w_j. $CONN_{ij}$ is calculated from a recall of the entire dataset as

$$CONN_{ij} = CADJ_{ij} + CADJ_{ji} \tag{5}$$

$$CADJ_{ij} = \sum_s I(BMU1(x_s) = i \wedge BMU2(x_s) = j), \tag{6}$$

where $BMU\{1, 2\}$ are the index of the 1st and 2nd **B**est **M**atching **U**nits (prototypes) and $I()$ is the indicator function. $CADJ_{ij}$ (the **C**umulative **ADJ**acency of i and j) reports the number of data vectors observed in the second-order Voronoi cell V_{ij} generated by W in $\mathbb{R}^d$, and CONN is its symmetrized version. CONN is thus a weighted version of the Masked Delaunay Triangulation [10,9] whose edge weights reflect local data densities and connectivities within the manifold. We note for later discussion that CONN is typically very sparse.

3. CONNt-SNE

CONNt-SNE methodology comprises two key modifications to t-SNE's definition of high-d similarity. The first permits a varying perplexity px_i when setting each conditional distribution $p_{\cdot|i}$ (recall
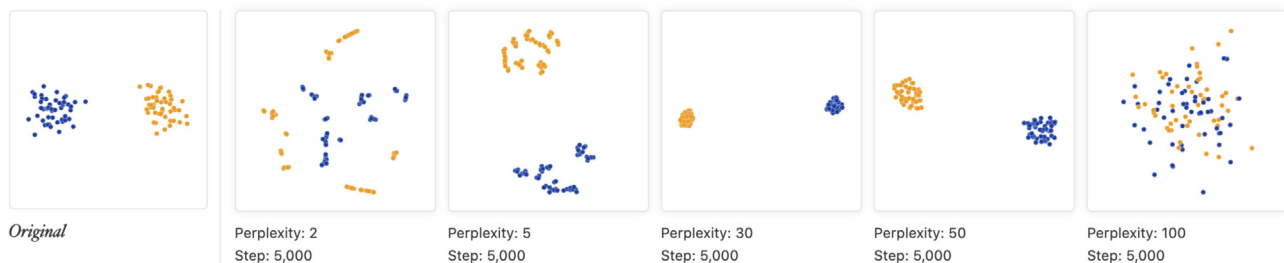

Original | Perplexity: 2 Step: 5,000 | Perplexity: 5 Step: 5,000 | Perplexity: 30 Step: 5,000 | Perplexity: 50 Step: 5,000 | Perplexity: 100 Step: 5,000

Fig. 1. The sensitivity of t-SNE embeddings to their parameterization, taken from [4].

from (2) that perplexity controls the Gaussian bandwidths σ_i which form the prototype similarities p_{ij}). We now have M different (local) perplexities to specify but CONN provides a data-driven way of determining these parameters as the number of CONN neighbors of prototype w_i, which we denote by v_i for the remainder of this work:

$$px_i = v_i = max\left(\sum_j I(CONN_{ij} > 0), 2\right). \tag{7}$$

It is possible that some prototype w_i has no CONN neighbors ($v_i = 0$), which occurs if a) the receptive field of i is empty and b) no datum has chosen i as BMU2. To avoid numerical issues we enforce a lower bound $px_i \geqslant 2$ in the above, but suggest removing such unused prototypes from W prior to running CONNt-SNE. With px_i intelligently and automatically specified, the same procedure of (2) sets each local σ_i (and, consequently, P_i). We denote by P_v the matrix of prototype similarities arising from CONN-derived variable perplexities px_i.

The second modification to t-SNE infuses the topological adjacency and local density information contained in the $CONN_{ij}$ values into the high-d similarity definition. This information can be viewed from two vantage points (scales). A *global* view (where each $CONN_{ij}$ value is considered relative to all other $CONN_{kl}$) grades the topological connectivities of major/coarse structures within the manifold, as learned by the vector quantizer. This information is most useful for characterizing regions of higher data density. We define a globally normalized version of CONN as

$$GCONN_{ij} = \frac{CONN_{ij}}{\sum_{kl} CONN_{kl}}, \tag{8}$$

and note that $\sum_{ij} GCONN_{ij} = 1$. In contrast, a *local* view (where each $CONN_{ij}$ is considered relative to all other $CONN_{i\cdot}$, i.e., when the CONN graph is viewed node by node) elicits finer structure in the manifold, particularly in areas of low data density. A locally normalized version of CONN is given by

$$LCONN_{ij}^* = \frac{CONN_{ij}}{\sum_k CONN_{ik}} \tag{9}$$

$$LCONN_{ij} = (LCONN_{ij}^* + LCONN_{ji}^*)/(2M), \tag{10}$$

where the last equation above is merely symmetrizing and re-normalizing LCONN∗ to have unity sum.

Ideally, we would like to imbue t-SNE with both (global & local) topological views offered by CONN, as these have been shown effective for inferring structure from complex manifolds such as hyperspectral imagery of Earth [11,12] and Mars [13], radioastronomy imagery [14], and functional MRI images of brains [15]. We achieve this multi-scale view by defining the following composite similarity to assess relationships in $\mathbb{R}^d$:

$$P_{CONN} = \frac{1}{3}(P_v + GCONN + LCONN). \tag{11}$$

The averaging of t-SNE's Gaussian-based similarity with the global and local views of manifold topology offered by CONN is similar in spirit to the multi-scale similarity proposed for stochastic neighbor embeddings in [16]. In that work, an aggregate high-d similarity is averaged from those resulting from an exponentially increasing set of perplexities in a range whose lower bound is user-specified and upper bound is data dependent. In contrast, CONNt-SNE utilizes an entirely different type of information in its multi-scale view, combining explicit notions of manifold connectivity and density, as expressed by CONN. We note that this type of information is unique to vector quantizers.

The attractive forces among embedded points in t-SNE are set by P (Eq. (1)) while the repulsive forces are governed by Q (Eq. (3)) [17]. Because CONN (and, consequently, GCONN and LCONN) is typically very sparse, use of P_{CONN} should cajole embedded points corresponding to adjacent prototypes closer together in the embedding. On the other hand, CONN dis-similarity of all non-adjacent prototypes i and j are the same, $CONN_{ij} = 0$. We have summarized the various similarity measures discussed in this work in Table 1.

As CONNt-SNE only modifies the high-dimensional similarity measure P used in the t-SNE framework it, by itself, imparts no additional computational effort to embed learned prototypes of data. However, obtaining these prototypes *does* require additional computation whose complexity obviously varies with the method by which the prototypes are obtained (e.g., online vs. batch versions of k-means, SOM, NG, etc.). We suggest that, regardless of the additional computation required, prototype-based learning is useful in its own right and is worth the additional time it requires. t-SNE has computational complexity $\mathscr{O}\left(N^2\right)$ when embedding N data points [2]; because prototype learning reduces sample size (often, the number of learned prototypes $M = \mathscr{O}\left(\sqrt{N}\right)$) immediate savings for t-SNE computational effort are achieved. Additionally, we point out that hardware implementations of neural prototype learning exist [18] making the prototype learning step near-instantaneous, even for large datasets.

4. Quality measures

While visual inspection of CONNt-SNE's embeddings is important, we have also assessed the quality of each embedding in two different quantitative categories: topology preservation, and the preservation of cluster structure as measured by Cluster Validity Indices. Throughout this work we indicate whether a larger or smaller value of a measure is preferred with up ↑ and down ↓ arrows, respectively.

4.1. Measuring topology preservation

K-ary neighborhood preservation [16], commonly used to assess the performance of dimensionality reduction techniques, measures the proportion of a high-d k-nearest neighborhood around each w_i that is preserved after embedding w_i in low-d, averaged over i. As this K-ary measure ([[16] Eq. 15]) yields a performance curve over $k \in \{1, \ldots, M-1\}$, we report the area under such curve (AUC), normalized by its theoretical maximum (M-2), for comparison across datasets.

K-ary neighborhood measures are one example of a family of topology preservation (TP) measures, but there are others. Drawing from the literature on Self-Organizing Maps we have also measured the mismatch between the topology of the manifold (which we call "input space") and its representation in the embedding (which we call "output space"), as reported by the Normalized Differential Topographic Function (NDTF [19], which is a differential form of the Topographic Function of [20], normalized to have unity sum). For a SOM, CONN (as a Topology Representing Network [21]) represents input space topology while the user-specified lattice defines the output space topology. In this work, CONN persists as a representation of input space topology, and we prescribe the output space topology as the Delaunay triangulation [22] of a t-SNE embedding $T \subset \mathbb{R}^2$, which we denote by $\mathscr{D}^T$. In what follows we also denote by $\Delta_{ij}^{\mathscr{D}^T}$ and Δ_{ij}^{CONN} the geodesic distance between prototypes i and j as measured on the $\mathscr{D}^T$ and CONN graphs, respectively.

The NDTF (and its relatives) all measure the degree to which the output space accurately reflects topological adjacencies in input

Table 1
An overview of point similarities considered in this work.

Similarity	Description
P, Q	Standard t-SNE point similarities (in the original space and embedded space, respectively) from [2].
CONN	The CONN similarity [9] expressing manifold connectivity between prototypes.
LCONN, GCONN	**L**ocally and **G**lobally normalized CONN similarities, see eqs. (8) and (10).
P_v	t-SNE's point similarity using a variable perplexity for each point, set by CONN. See eq. (7).
P_{CONN}	The similarity used for CONNt-SNE, aggregated from P_v, GCONN & LCONN.

space (a measure of *forward* topology preservation), and vice versa (a measure of *backward* topology preservation). Specifically, the forward measure,

$$\text{FNDTF}(r) = \hat{E}\left[I\left(\Delta_{ij}^{\mathscr{D}^T} = r\right)|\Delta_{ij}^{\text{CONN}} = 1\right],$$

reports the proportion of prototype adjacencies on CONN that are of geodesic distance r on $\mathscr{D}^T$. Similarly, the backward measure

$$\text{BNDTF}(r) = \hat{E}\left[I\left(\Delta_{ij}^{\text{CONN}} = r\right)|\Delta_{ij}^{\mathscr{D}^T} = 1\right]$$

reports the proportion of prototype adjacencies on $\mathscr{D}^T$ that are of geodesic distance r on the CONN graph. In the above, $\hat{E}$ denotes the conditional empirical mean over all relevant adjacencies ij and $I(A)$ is the indicator function of event A. We note that the Forward and Backward (abbreviated F/B here) NDTFs both have unity sum over r, and that NDTF(0) is undefined. Typically, an analyst would view the trace plot of (F/B) NDTF(r) vs. r to assess the exact location (geodesic distance r) and severity (the value (F/B) NDTF(r)) of observed topology violations. As $r = 1$ is not considered a violation, (F/B) NDTF(1) = 1 conveys perfect topology preservation while any (F/B) NDTF($r > 1$) > 0 indicates violations. In order to combine both the location and severity of topology violations into one measure we define the **F**orward/**B**ackward **T**opological **N**eighborhood **E**xpansion as:

$$(\text{F/B})\text{TNE} = \sum_{r>0} r \times (\text{F/B})\text{NDTF}(r). \tag{12}$$

(F/B) TNE reports the average geodesic radius by which a topological neighborhood in one space (CONN/$\mathscr{D}^T$) must expand to be represented in another space ($\mathscr{D}^T$/CONN). A perfect embedding by this measure has (F/B) TNE = 1, and topology violations of increasing severity are reported by values > 1.

4.2. Measuring structural preservation with cluster validity indices

Faithful topology representation is desirable when inferring (cluster) structure of a high-d manifold from its embedding. While *sufficient*, exact TP may not be *necessary* for structural identification; indeed, according to the TPMs discussed above, the rigid 2-d SOM lattice cannot faithfully represent manifold topologies with more than a few (8 for rectangular lattices, 6 for hexagonal) neighbors, but this fact has not hindered its success as a tool for cluster discovery. To account for this we have measured the structural preservation of our experimental embeddings, as reported by a variety of internal and external Cluster Validity Indices (CVIs). As t-SNE is most commonly used to identify such structure (or lack thereof), we believe these measures better reflect the quality of an embedding for most uses of t-SNE in practice (See Table 2).

4.2.1. Internal cluster validity indices

Internal CVIs (CVIIs) measure the relationship between compactness and separation of clusters defined by a given partitioning of the data. While there are many such measures [23], in this work we focus on three of the more common: the (average) Silhouette Index ↑ (SIL, [24]), Generalized Dunn Index ↑ (GDI, with set distance 5 and set diameter 3, as defined in [25]), and Davies-Bouldin Index ↓ (DBI, [26]). As our goal is to show how well cluster structure is preserved when embedding high-d prototypes $W \subset \mathbb{R}^d$ as $T \subset \mathbb{R}^2$, we report CVIIs measured on the latter relative to those measured on the former, using the true partitioning of each dataset ℓ^*. For example, the relative Silhouette of an embedding, $\text{rSIL}_W^T = [\text{SIL}(T) - \text{SIL}(W)]/|\text{SIL}(W)|$, measures the change in the Silhouette score of the true partitioning after embedding W by T, relative to its value in W. rGDI and rDBI are computed similarly. As there is no universally best CVII in all cases, we average the individual relative CVIIs into an aggregate score:

$$\uparrow \text{rCVII}_W^T = \frac{1}{3}(\text{rSIL}_W^T + \text{rGDI}_W^T - \text{rDBI}_W^T), \tag{13}$$

where the subtraction above arises because a lower value of DBI is preferred. Whereas the K-ary score reports preservation of Euclidean distances and the (F/B) TNE scores report preservation of topological distances, rCVII signals preservation of set cohesion and distance, where the sets are the clusters of the true data partitioning. Thus, rCVII should give some indication of how the embedding (mis)represents cluster structure which, in turn, hints at its impact on 2-d cluster inference.

4.2.2. External cluster validity indices

Both to assess whether rCVII is performing as designed, and to simulate how actual clusterings are affected by the process of embedding, we also cluster $W \subset \mathbb{R}^d$ and $T \subset \mathbb{R}^2$, resulting in partitionings ℓ^W and ℓ^T, respectively. The quality of each partitioning is assessed, relative to the truth ℓ^*, by several External CVIs (CVIEs): the Adjusted Rand Index ↑ (ARI, [27]), Jaccard Index ↑ (JAC, [28]), and Normalized Mutual Information ↑ (NMI, [29]). Again, we are more interested in comparing the quality of ℓ^T relative to ℓ^W, rather than the absolute value of either, which we accomplish via the relative measure $\text{rARI}_W^T = [\text{ARI}(\ell^T, \ell^*) - \text{ARI}(\ell^W, \ell^*)]/|\text{ARI}(\ell^W, \ell^*)|$ (and similarly for JAC and NMI). An aggregate measure of relative external cluster validity is defined as

$$\uparrow \text{rCVIE}_W^T = \frac{1}{3}(\text{rARI}_W^T + \text{rJAC}_W^T + \text{rNMI}_W^T). \tag{14}$$

Computing rCVIE obviously requires a clustering, which we obtain via Spectral Clustering with random walk normalization of the graph Laplacian [30], as there is a purported connection between certain parameterizations of t-SNE and spectral clustering [31]. The true number of data clusters (C from Table 3) parameterize the k-means step of the spectral clustering procedure.

5. Experiment design

5.1. Datasets

To demonstrate the effectiveness of CONNt-SNE we compare its two-dimensional embeddings to those of t-SNE for the six real datasets (indexed by δ) whose characteristics are given in Table 3. These include Standard COIL20 [32] (labeled as in Fig. 2) and MNIST [33] along with two of MNIST's more challenging drop-in replacements: Fashion MNIST (FMNIST, [34]), containing images of 10 different articles of clothing, and Kazushiji MNIST (KMNIST, [35]), containing images of 10 different Japanese Hiragana characters. Both MNIST replacements have 28×28 pixel images. The

Flow18 dataset contains flow cytometry measurements of 946,915 human peripheral blood mononuclear cells labeled by 12 different phenotypes, subsampled as in [36] (we have ignored the "Dead cells" class in this analysis). Ocean City (OC) is a 512×512 pixel, 8-band spectral image of Ocean City, Maryland, with 1.5 m/pixel resolution. Data collection, pre-processing and mean signatures of verified land-cover classes are given in [11]. We consider the 29 clusters interactively identified in [9] as truth clusters. These clusters comprise three larger material groupings—vegetation, water, man-made materials—each broken down into a number of unique clusters with widely varying statistical properties (see representative statistics in [37]).

5.2. Experiment descriptions

The various t-SNE methods and their nomenclature utilized in this work are presented in Table 4. To alleviate notation we will use a method's name (e.g., t-SNE(10)) and its similarity (e.g., P_{10}) interchangeably, as the similarity uniquely defines the method. Thus we have 7 methods $h \in \{P_{10}, P_{20}, P_{30}, P_{40}, P_{50}, P_{MS}, P_{CONN}\}$, where P_{MS} is the multi-scale method of [16].

To make our conclusions more robust, for each dataset and method we have produced 200 different embeddings resulting from initializations $\iota \in \{PCA, 1, \ldots, 199\}$, where PCA denotes a 2-$d$ principal components initialization, and integers 1–199 represent a randomly seeded initial state. Thus, there are 200 embeddings for dataset δ using method h, and T_h should be viewed as a function $T_h(\delta, \iota)$. 6 datasets $\times$ 7 methods $\times$ 200 initial states yields 8,400 embeddings from which we draw conclusions.

We assess these 8,400 embeddings with the five quality measures $\mu \in \{FTNE, BTNE, K\text{-}aryAUC, rCVII, rCVIE\}$ described in Section 4. However, each μ measures different characteristics of our embeddings and, consequently, possesses a wide range of scales; this complicates comparison amongst the μ, and across different datasets. To facilitate such meta-analysis we will report instead a standard score $Z_\mu(T)$ using $\mu(T_{30})$ as a baseline, as $px = 30$ is a widely used default in popular t-SNE implementations. Thus, for each measure μ of each embedding of each dataset, $T_h(\delta, \iota)$, we report

$$Z_\mu(h, \delta, \iota) = \frac{\mu(T_h(\delta, \iota)) - \mu(T_{30}(\delta, \iota))}{\hat{\sigma}[\mu(T_h(\delta, \cdot)) - \mu(T_{30}(\delta, \cdot))]},$$

where $\hat{\sigma}$ is the empirical standard deviation of the measure differences, computed over the 200 different initializations ι. Additionally, for consistency, we report $-Z_{FTNE}$ and $-Z_{BTNE}$, as lower values of these measures indicate better topology preservation. This makes all Z-scores comparable.

Not only is $Z_\mu(h, \delta, \iota)$ unitless, its mean is: 1) the effect size of method h relative to P_{30} (also known as Cohen's d [38]), and 2) proportional to the test statistic of a paired Student's t-test of the above. Thus, differences in summary statistics of $Z_\mu(h, \delta, \iota)$ immediately reveal statistical significances of the effect sizes of each method. Since P_{30} is used as a basis for standardization, its Z-scores are all = 0, and will be excluded from visualization of results.

All results that follow were produced by minimizing t-SNE's cost function (4) with Delta-Bar-Delta gradient descent (as in [2]) for a maximum of 2,000 iterations, monitored every 50 iterations. Early stopping was permitted if the cost function decreased by $<0.1\%$ for 3 consecutive monitoring steps (150 learning steps). The learning rate for gradient descent was set = 200, with momentum increased from 0.5 to 0.8, in line with [2]. Although the use of exaggeration (inflating the high-d similarities P by some constant α) is widely thought to improve the minimization of (4) and avoid crowding in the embedded space, no consensus on how much exaggeration to use, or how long to enforce it, seems to exist. Various contradictory work recommends both early and late scheduling of high and low values of α [2,36,39,31,40], while a preprint suggests use of exaggeration may fundamentally alter the nature of t-SNE altogether [17]. As a result, we have taken the conservative recommendation of [17] and linearly annealed α from 4 to 2 over the 2,000 prescribed learning steps to effect mild versions of both early and late exaggeration schemes. Batch neural gas learning [7] generated the prototypes used in this work for all datasets except Ocean City, where the previously scrutinized SOM prototypes from [9] were used for comparative consistency.

6. Results

6.1. Meta-analysis I: overall aggregated results

Fig. 3 reports standardized effects $Z_\mu(T)$ for each method, showing overall aggregated effect sizes (top panel, [a]) and those aggregated by dataset (middle panel, [b]) and measure (bottom panel, [c]). The Z_μ scores for individual measures were combined in pan-

---

Table 2
An overview of various measures used to assess the quality of t-SNE embeddings in this work. Up (down) arrows indicate a larger (smaller) value of the measure is preferred.

Topology Preserving Measures	
K-ary ↑	The K-ary neighborhood preservation score, conveying the proportion of k-nearest neighbors that are preserved in the high- and low-dimensional spaces [16]
(F/B) NDTF (r)	The Forward/Backward **N**ormalized **D**ifferential **T**opographic **F**unction [19], measuring the proportion of topologically adjacent neighbors in one space (e.g., high-d) that are of geodesic distance r in another space (e.g., low-d embedding), according to its topology. FNDTF($r = 1$) = 1 indicates the topology of high-d space is perfectly represented by the topology of the embedded space, while BNDTF($r = 1$) = 1 indicates the opposite. Any (F/B) NDTF($r \geqslant 2$) > 0 indicates topological mismatch (folding) between the two spaces under consideration.
(F/B) TNE ↓	The Forward/Backward **T**opological **N**eighborhood **E**xpansion, indicating the average neighborhood size (geodesic radius) in one space that is required to represent the immediate topological neighbors of another space. Newly presented in this work, TNE aggregates the NDTF over all possible geodesic distances r, combining the severity and location of topological violations into one measure.

Cluster Validity Indices	
rCVII ↑	Relative Internal Cluster Validity Indices (13), combining the quality of a clustering as reported by three widely used internal measures: the Silhouette Index [24], the Generalized Dunn Index [25], and the Davies-Bouldin Index [26]. Here, "relative" indicates that the CVIIs values computed in the low-d embedding are reported relative to their measure in high-d space. This relative measure overcomes difficulties when comparing assessments of cluster quality based on distances in two different spaces.
rCVIE ↑	Relative External Cluster Validity Indices (14), combining the agreement of a clustering to a true (known) partitioning as reported by the Adjusted Rand Index [27], the Jaccard Index [28], and Normalized Mutual Information [29]. Here, "relative" indicates that the reported values are measures of CVIEs resulting from a clustering of the low-d embedding, reported relative to those obtained from a clustering in high-d space. The clusterings in each space were obtained by Spectral Clustering [30], as [31] suggests there is a connection between it and t-SNE.

Table 3

Characteristics of the six datasets used in experiments: sample size (N) and dimension (d), the number of sample classes (C), the number of prototypes which learned the data (M), and a 95% confidence interval for the average number of CONN neighbors ($\bar{v}$). Points from the "Dead cells" class of Flow18 were removed prior to our analysis, along with unlabeled pixels in the Ocean City image. M reported in this table excludes any unused prototypes.

Data	N	d	C	M	$\bar{v}$
MNIST	70,000	784	10	2,000	$_{18.9}19.2^{19.5}$
FMNIST	70,000	784	10	2,000	$_{14.6}14.8^{15.1}$
KMNIST	70,000	784	10	2,000	$_{16.9}17.3^{17.7}$
COIL20	1,440	16,384	20	492	$_{2.3}2.3^{2.4}$
Flow18	946,915	11	12	1,457	$_{30.3}30.8^{31.4}$
OC	251,946	8	29	1,464	$_{11.4}11.6^{11.8}$

Fig. 2. COIL20 image database with integer encoded labels.

els [a] and [b], since they are now comparable. Violin plots show the distribution of effect sizes by method, with black error bars displaying the estimated mean (with 95% confidence interval) of each. Purple points report the scores of the PCA-initialized embedding separately, as informative (non-random) initializations are recommended in [41]. The green lines at $Z = 0$ represent the P$_{30}$ case serving as baseline, and green numbers report the proportion of experiments for each particular method which induce a positive effect (i.e., the estimated probability $Pr[Z > 0]$). Detailed statistics by each measure and dataset are in Fig. 6.

P$_{CONN}$ induces the largest positive average effect (0.78, annotated in black numbers for clarity) over all experiments, as reported in Fig. 3[a]. Although it is hard to detect from the confidence intervals shown at this scale, the overall effect of P$_{CONN}$ is statistically larger (at significance level $\alpha = 0.05$) than both P$_{10}$ and P$_{20}$, which jointly performed second-best (their performance is statistically indistinguishable, $\alpha = 0.05$). Aggregate performance of regular t-SNE degrades monotonically as px increases, although this may be an over-generalization (addressed below). The mean

performance trends are also supported by non-parametric statistical arguments, where the proportion $Pr[Z > 0]$ is estimated at 0.53, 0.51, and 0.54 for P$_{CONN}$, P$_{10}$ and P$_{20}$, respectively. Binomial tests ($\alpha = 0.05$) of these proportions reveal P$_{CONN}$ and P$_{20}$ result in measurable improvements to P$_{30}$ embeddings more than half the time (a similar test for P$_{10}$ produced a p-value = 0.09). P$_{CONN}$'s Z-distribution exhibits pronounced positive bimodality and skew, while both mode location and skew appear negatively correlated with perplexity in regular t-SNE. Overall from Fig. 3[a] we conclude that, for these experimental data, P$_{CONN}$, P$_{10}$ and P$_{20}$ all produce reliable improvements to the P$_{30}$ baseline, with P$_{CONN}$'s mean effect size (0.78) more than twice as large as P$_{10}$/P$_{20}$ (0.37). For completeness we note that all three information streams comprising P$_{CONN}$ (P$_v$, GCONN, and LCONN (11)) induced positive effects, but their combination is best.

6.2. Meta-analysis II: results by dataset

Fig. 3[b] reveals most, but not all, datasets obey the generalization that performance of regular t-SNE deteriorates monotonically with perplexity, which is not surprising given the large variation in sample size (here, number of prototypes), structural complexity, and inherent dimensionality of the data considered in this work. For example, COIL20 exhibits a statistically significant effect size improvement from P$_{20}$ to P$_{50}$, and t-SNE for Ocean City has largest average effect at P$_{40}$. Interestingly, OC is also the only dataset for which CONNt-SNE's *mean* effect size is neither best, nor statistically positive (although its PCA initialized case *is* still superior to its counterparts). This is likely due to the level of noise in Ocean City's spectra, as well as the spectral similarity of its 29 known clusters (which are subclusters of three large material tranches: vegetation, water, and man-made materials). Because of this, [[9] Section 4B] removes $\sim 20\%$ of Ocean City's CONN edges, according to a thresholding scheme defined therein, to facilitate clustering. We believe some degree of CONN edge removal would also benefit CONNt-SNE, but have left this for future work. Despite this, Section 6.4 discusses the visual improvement of P$_{CONN}$'s OC embedding, compared to P$_{20}$.

Table 4

Nomenclature for the methods under comparison. var + indicates CONNt-SNE utilizing the topological information in CONN, in addition to a variable perplexity.

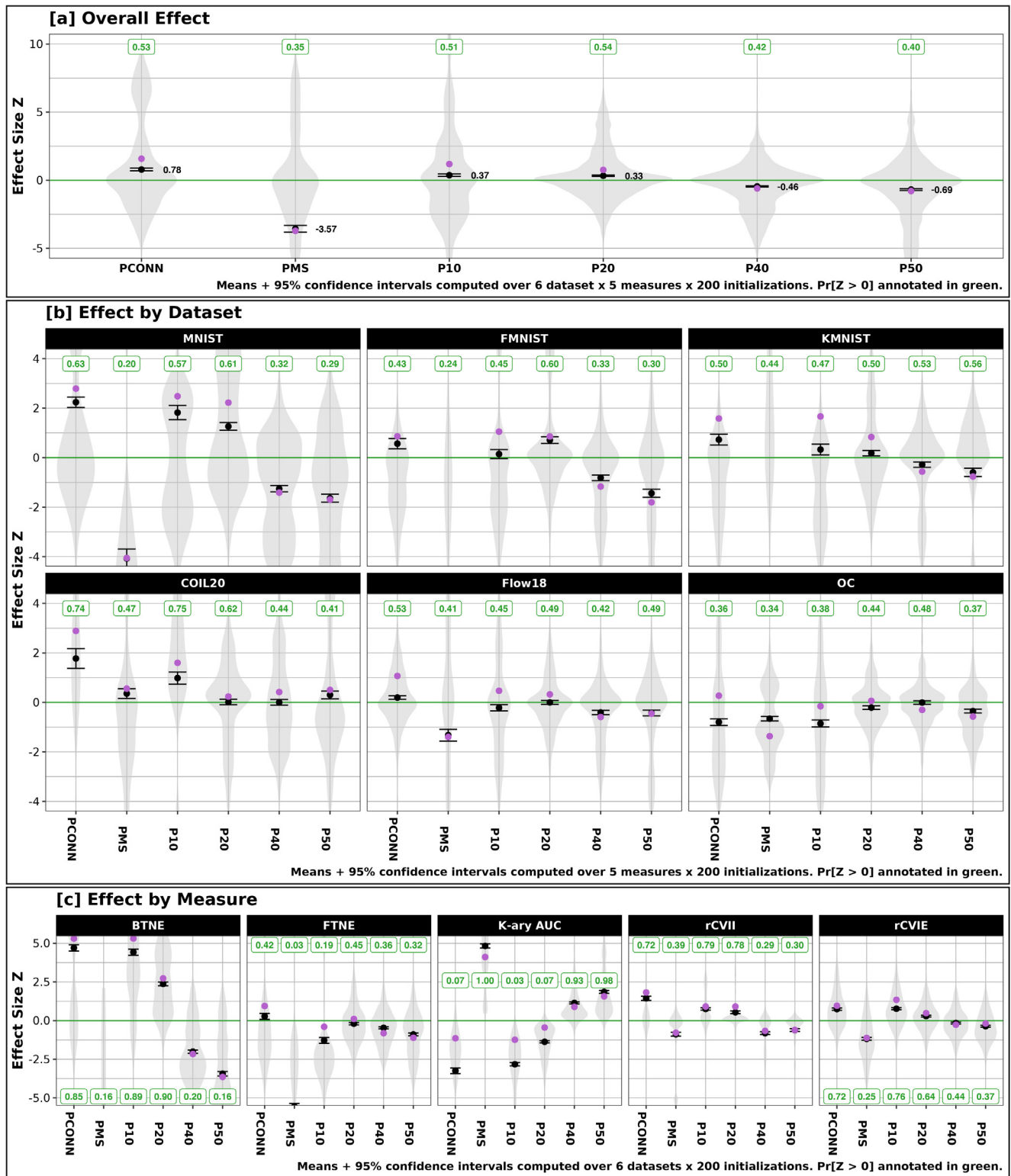| Method	Perplexity	Similarity	Embedding
t-SNE(10)	10	P$_{10}$	T$_{10}$
t-SNE(20)	20	P$_{20}$	T$_{20}$
t-SNE(30)	30	P$_{30}$	T$_{30}$
t-SNE(40)	40	P$_{40}$	T$_{40}$
t-SNE(50)	50	P$_{50}$	T$_{50}$
MS t-SNE	var	P$_{MS}$	T$_{MS}$
CONNt-SNE	var+	P$_{CONN}$	T$_{CONN}$

Fig. 3. Performance of each method, according to the standardized score of each measure. Panel [a] reports an overall aggregation by method whereas [b] and [c] aggregate performance by dataset and measure, respectively. Error bars convey means and 95% confidence intervals over 1,400 total experiments (200 different initializations × 6 datasets) with each method, while the gray shaded violin plots show the entire distribution of values underpinning these error bars. Separate purple points represent the measure of the PCA-initialized embedding. Green lines at $Z = 0$ represent the performance of its baseline P_{30} (see Section 4), and green numbers report the estimated probabilities $Pr[Z > 0]$, which convey the proportion of time each method has positive effect.
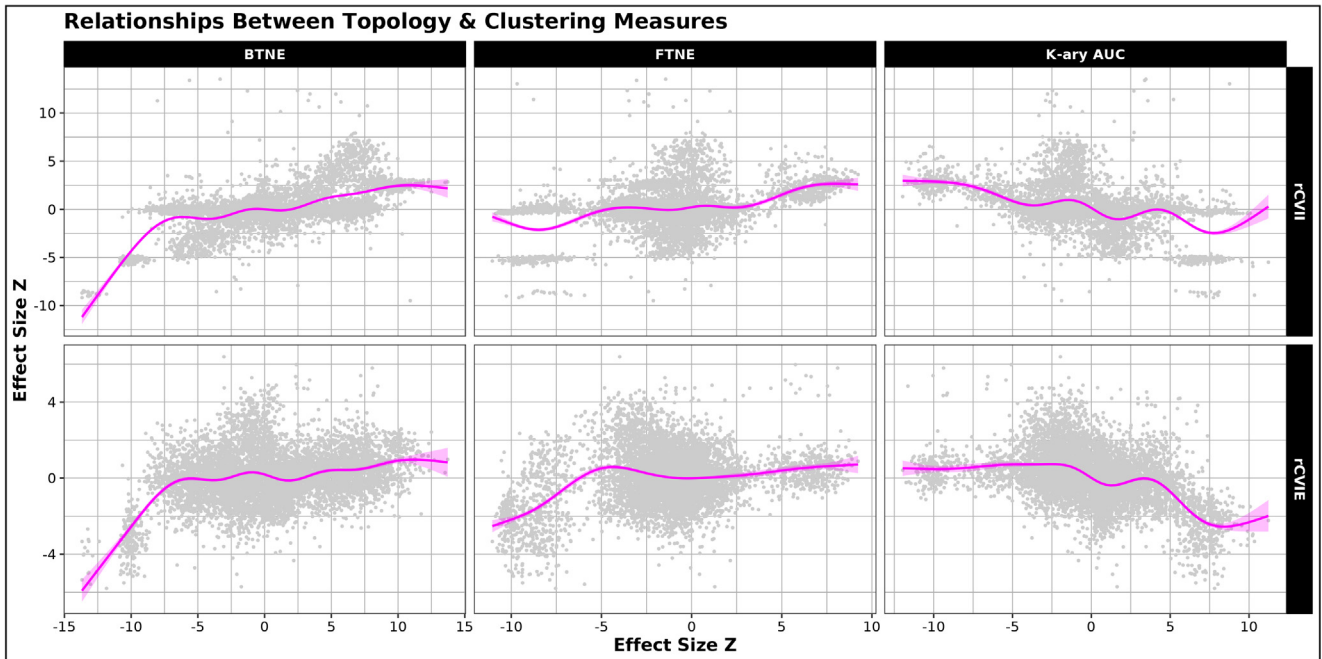
Fig. 4. Relationships of Cluster Validity Indices to different Topology Preservation Measures for the 8,400 embeddings studied in this work (gray points). Pink banded trendlines in each panel report 95% predictive intervals of a spline regression, fit via a Generalized Additive Model [42] which selects the level of model smoothing automatically.

6.3. Meta-analysis III: results by measure

From Fig. 3[c], P_{CONN} is the only method with a statistically positive mean effect size by all measures, excluding the K-ary score. We expect it to achieve higher BTNE and FTNE measures, as the GCONN and LCONN components increase the similarity p_{ij} of topological neighbors i and j in P_{CONN}. This influence appears to help CONNt-SNE preserve cluster structure better, resulting in higher rCVII scores, which are positively correlated to rCVIE scores (0.49±0.02 overall, at 95% confidence). Positive rCVIE effects show that, when properly parameterized, t-SNE can be an effective tool for feature engineering. Recall that rCVIE reports relative change in external CVI measures of a partitioning obtained via spectral clustering of the embedded points, versus one obtained by clustering the prototypes in $\mathbb{R}^d$. We have employed spectral clustering (as a widely accepted and trusted clustering method) in this work and acknowledge that other clustering regimes may impart different effects. However, this analysis does support further exploration of t-SNE as a pre-processing step in larger machine learning pipelines, particularly where linear pre-processing (e.g., PCA) is inappropriate.

Our discussion of results up to this point has ignored the performance of the multi-scale similarity P_{MS}, which is lowest in overall aggregate. An explanation for these low scores is found in Fig. 3[c], which shows P_{MS} fails to produce either mean or median positive effects according to the (F/B) TNE and CVI measures. P_{MS} does, however, achieve significantly higher K-ary scores than all other methods. This agrees with the conclusions presented in [16], where MS similarities were shown to increase K-ary scores for a variety of dimension reduction algorithms, including Stochastic Neighbor Embedding. The P_{MS} similarity is obtained by averaging P_{px} over exponentially increasing px bound by a range intended to be large enough to enforce global ordering, and small enough to avoid uniformity of its values. In this work we set the lower bound = 10 (the same used for the px grid); the upper bound is data-dependent (set as in [[16] Section 3.1]) but is generally much

higher than our px grid upper bound of 50 (e.g., for MNIST with 2000 prototypes, P_{MS} is influenced by px in a range from 10 to 1280), which directly correlates with its ability to preserve Euclidean neighborhood ordering across a large range of neighborhood sizes.

But is this a desirable characteristic of an embedding in practice? That is, which neighborhoods should be preserved to most faithfully represent cluster structure in an embedding? As structure identification motivates most uses of t-SNE we have explored this question a bit further. Fig. 4 displays scatterplots of the Z-scores (gray points) of each CVI vs. each topology preservation measure considered in this work. A non-linear spline regression with corresponding 95% predictive interval is shown as a pink banded trend line. Here, the regression was fit via a Generalized Additive Model [42] which automatically (jointly) optimizes the level of smoothing. The CVI vs. (F/B) TNE trends are statistically significant (p-value $\approx$ 0) and positive (i.e., better topological neighborhood preservation is associated with better clustering results). CVI vs. K-ary trends are also significant (p-value $\approx$ 0) but generally negative overall. Thus, a high K-ary score appears inversely (or, at least not positively) related to t-SNE's preservation of cluster structure. Stated another way, demanding full Euclidean neighborhood preservation from an embedding algorithm may be intuitively desirable, but appears an overly conservative constraint. This is in line with the literature on CONN-based clustering [12,37] which concludes that topological characterizations of locality are more beneficial than their Euclidean analogs for extracting structure from data.

6.4. Visual inspections

In closing, we discuss some qualitative aspects of the embeddings visible in Fig. 5. For each dataset, CONNt-SNE embeddings are shown vertically atop the best performing case (according to Fig. 3[b]) of regular t-SNE, which is P_{10} for all except Ocean City (P_{20}). PCA initialized results are shown, as these outperformed
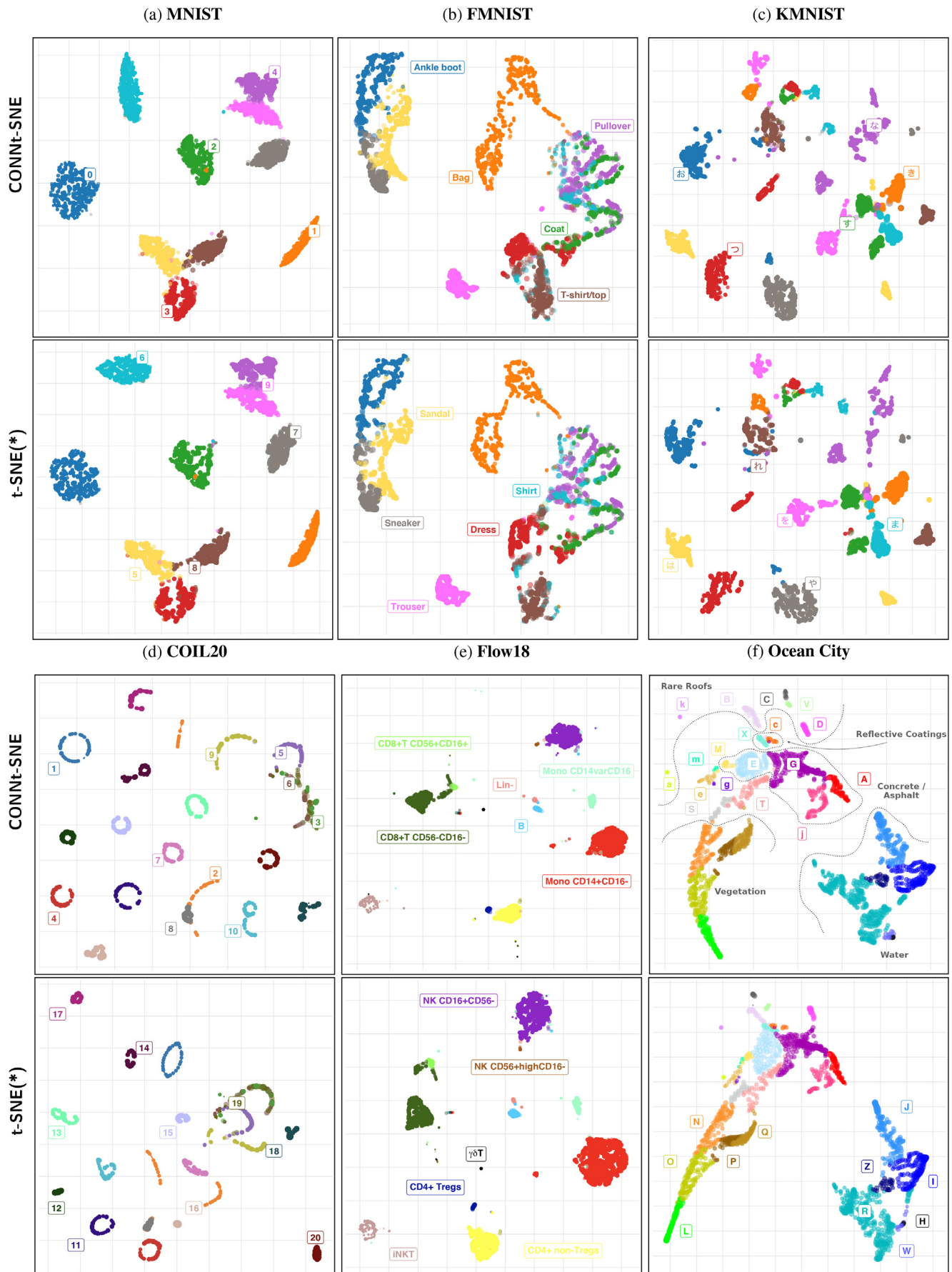
(a) **MNIST** (b) **FMNIST** (c) **KMNIST**

(d) **COIL20** (e) **Flow18** (f) **Ocean City**

Fig. 5. Embeddings of the prototypes of our experimental datasets. t-SNE(*) means P_{10} in all cases except Ocean City where P_{20} is shown. Prototype colors represent their learned truth labels. Label annotations are split across the pair for space considerations.
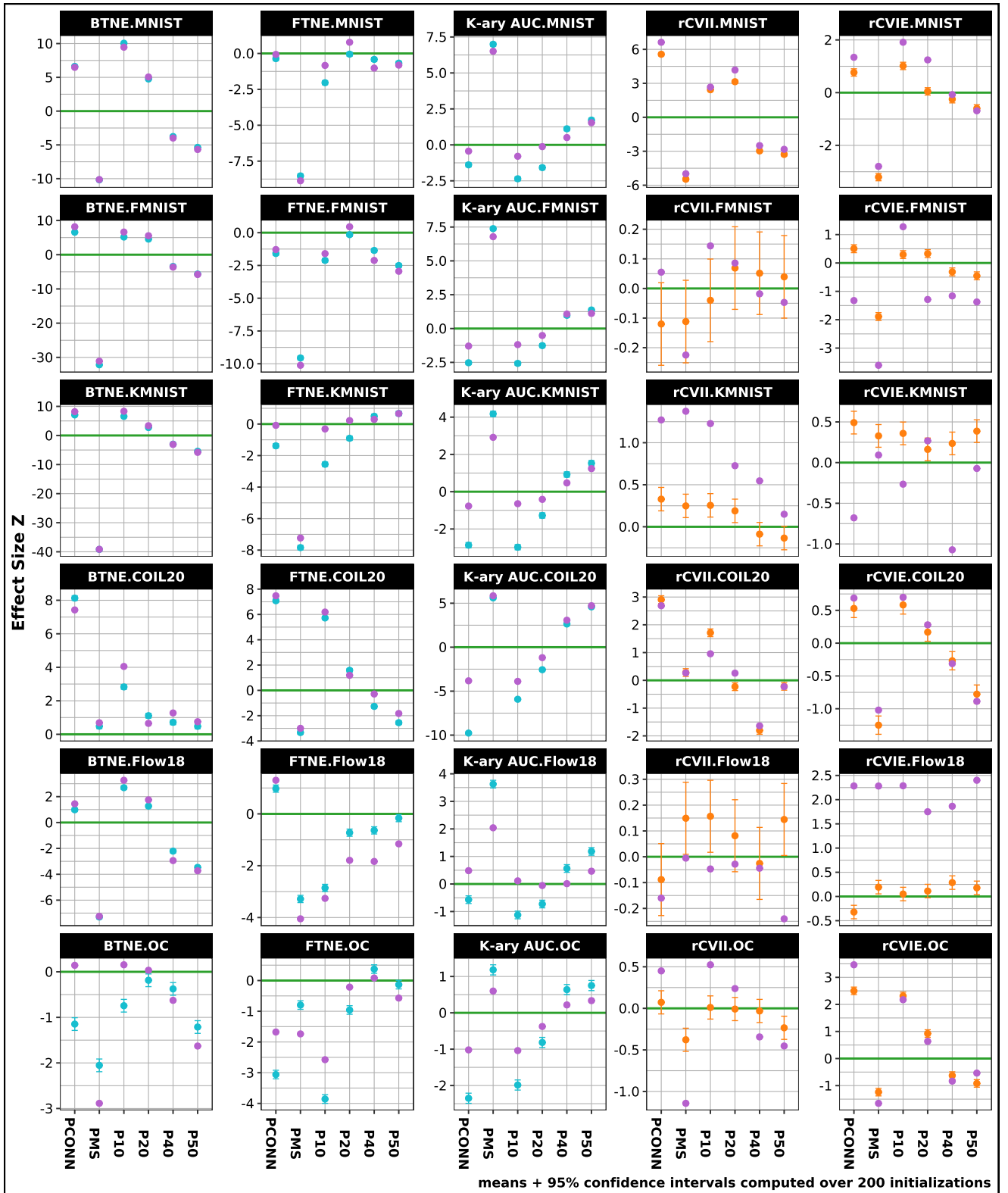
Fig. 6. The performance measures considered in this work (summarized in Table 2) for each experimental dataset (summarized in Table 3). These measures underpin the aggregate performances depicted in Fig. 3, and are shown here for completeness.

most randomly initialized embeddings according to Fig. 3. Point colors represent the prototype's true class label (decided via plurality vote of its receptive field), and point sizes represent the (relative) size of the prototype's receptive field. The annotations in each panel describe the true class labels (represented by different colors) for each dataset: for MNIST these are handwritten digits 1–10; for FMNIST they are articles of clothing; for KMNIST the Hiragana characters are shown; the mapping for COIL20's integer encodings is shown in Fig. 2; for Flow18 the cell types from [36] are annotated; the different material types in the Ocean City spec-

tral image (e.g., concrete, waters, vegetation) are encoded by the same letters A,B,etc. as described in [9].

Overall, CONNt-SNE embeddings are very similar to those induced by the best performing regular t-SNE (P_{10}/P_{20}) which supports the main assertion of this work: CONNt-SNE's data-driven modifications to t-SNE's similarity cause no degradation to the quality of embeddings; in some cases they result in visual improvement. In what follows we point out a few details in the embeddings of each dataset for further discussion.

MNIST (panel Fig. 5(a)) shows cohesion of the digit clusters, with the easily distinguishable digits (0, 1, 2, 6) well separated; both P_{CONN} and P_{10} have delineated the components of the 4–7–9 and 3–5–8 digit super clusters, which are typically harder to embed. Fashion MNIST (panel Fig. 5(b)) is a bit more challenging. Both P_{CONN} and P_{10} have isolated the trouser and bag clusters well, along with a footwear super cluster which shows sensible internal arrangement. In contrast, the super cluster containing Pullovers, Shirts and Coats is very mixed. P_{10} has better separated the T-shirt cluster at the expense of also splitting the Dresses. P_{CONN} and P_{10} have both responded to the high intra-class variation in KMNIST (panel Fig. 5(c)) by creating several subgroupings of each class, which is more organized in some cases (e.g., the purple subclusters are at least near each other) than others (e.g, the pink and brown subclusters are not geographically close). KMNIST may not be separable in 2-*d*, as others have also reported poor results from a variety of dimension reduction techniques [43]. P_{CONN} has produced a visually superior COIL20 embedding (panel Fig. 5(d)), retaining better separation, and more of the known ring-like structure, of COIL20's classes than P_{10}. Likewise, P_{CONN} has maintained the integrity of the dark green CD8 + T cell cluster in Flow18 (panel Fig. 5(e)), but overall both P_{CONN} and P_{10} have produced embeddings visually superior to those previously published [36], [Fig. 1(b)].

From panel Fig. 5(f) we see that both P_{CONN} and P_{20} embed the subclusters of the larger material classes—water, vegetation, manmade materials—together. There is also meaningful organization within these superclusters. For example, the vegetation group has been ordered (when viewing the "tail" of the embeddings from bottom to top) by bright green (cluster *L*), pea green (*O*), then orange (*N*). These represent healthy green vegetation, yellow lawns, and dry grasses, respectively. The gray and salmon colored clusters (*S* and *T*) represent, respectively, bare soil and boat docks (dry woody material possibly mixed with concrete). Thus, it is sensible that *S* and *T* form a "bridge" between the vegetation and man-made material super clusters; further, there is more organizational meaning to the fact that the *S*/*T* bridge terminates at vegetation cluster *N* than, say, cluster *L* (the wood comprising the docks is more similar to bare soil and dry grass than to green vegetation, evident from the visible near-infrared spectral signatures of the classes shown in [9,11]). However, P_{CONN} elucidates a few interesting structural components of the Ocean City spectra that P_{20} misses. Of note, P_{CONN} fully separates cluster *P*/*Q* (brown) which represents muddy marshy land with spectrally similar (but still distinct) vegetation to the dry grass in orange cluster *N*. Similarly, P_{CONN} is more sensitive to the distinction among various man-made materials (e.g. clusters *X*/*c*) which P_{20} fails to fully distinguish. The cluster distinctions expressed by P_{CONN}'s OC embedding better agree with clusters found earlier [9,11].

7. Conclusions and future directions

We have presented CONNt-SNE as a data-driven alternative to cumbersome and tedious exhaustive grid searches for optimal t-SNE perplexity. CONNt-SNE relies upon, and benefits from, prototype representations of data, which 1) increase the speed and fea-

sibility of embedding large datasets with t-SNE (recall, $M \ll N$) and 2) offer unique views of data topology in the form of the CONN graph.

As a weighted version of the Masked Delaunay Triangulation [21], CONN [9] reports topological connectedness and separation across a manifold; we incorporate this information into automated specification of variable t-SNE perplexities for each prototype. We further sensitize t-SNE's high-*d* similarity to the *strength* of manifold connectivities, as reported by CONN's edge weights viewed at various resolutions (global, local). Both modifications are crucial to CONNt-SNE's performance which, as shown by experiments, meets or exceeds the best offered by regular t-SNE with grid-optimized perplexity.

We have also explored the relationship between K-ary neighborhood preservation, which is a popular quality measure of dimension reduction techniques, and the preservation of known high-*d* cluster structure in low-*d* embeddings. Experiments show high K-ary neighborhood scores do not necessarily translate to embeddings of highest fidelity to such structure. Manifold topology matters, both when assessing the quality of an embedding and when inferring structure from it. CONNt-SNE's data-driven ability to recognize and respond to structural subtleties in real data facilitates more confident and meaningful inference from its embeddings.

As CONNt-SNE is new we have many ideas for further work, including: extensions of its framework to other dimensionality reduction techniques, permitting embedding of out-of-sample data points through clever use of the VQ mapping, and sensitization of t-SNE's repulsive forces (*Q*) to learned manifold *dis*-connectedness, possibly by considering geodesic distances along the CONN graph.

8. Acknowledgements

Declaration of Competing Interest

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

References

[1] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, first ed., Springer Publishing Company, Incorporated, 2007.

[2] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[3] K. Taşdemir, E. Merényi, SOM-based topology visualisation for interactive analysis of high-dimensional large datasets, Mach. Learn. Rep. 1 (2012) 13–15.

[4] M. Wattenberg, F. Viégas, I. Johnson, How to use t-sne effectively, Distill (2016), https://doi.org/10.23915/distill.00002.

[5] T. Kohonen, Self-Organizing Maps, Springer, 2000.

[6] T.M. Martinetz, K.J. Schulten, in: T. Kohonen, K. Mäkisara, O. Simula, J. Kangas (Eds.), A "neural gas" network learns topologies: Proceedings of the International Conference on Artificial Neural Networks 1991, North-Holland, (Espoo, Finland), Amsterdam; New York, 1991, pp. 397–402.

[7] M. Cottrell, B. Hammer, A. Hasenfuß, T. Villmann, Batch and median neural gas, Neural Networks 19 (6–7) (2006) 762–771.

[8] K. Taşdemir, Dimensionality reduction based similarity visualization for neural gas, in, 2014 IEEE International Conference on Data Mining Workshop (2014) 668–675, https://doi.org/10.1109/ICDMW.2014.42.

[9] K. Taşdemir, E. Merényi, Exploiting data topology in visualization and clustering of self-organizing maps, IEEE Trans. Neural Networks 20 (4) (2009) 549–562, https://doi.org/10.1109/TNN.2008.2005409.

[10] T. Martinetz, K. Schulten, Topology representing networks, Neural Networks 7 (3) (1994) 507–522, https://doi.org/10.1016/0893-6080(94)90109-0.

[11] E. Merényi, B. Csathó, K. Taşdemir, Knowledge discovery in urban environments from fused multi-dimensional imagery, in: 2007 Urban Remote Sensing Joint Event, IEEE, 2007, pp. 1–13.

[12] E. Merényi, J. Taylor, Empowering graph segmentation methods with SOMs and CONN similarity for clustering large and complex data, Neural Comput. Appl. 32 (24) (2020) 18161–18178.

[13] K. Taşdemir, E. Merényi, Cluster analysis in remote sensing spectral imagery through graph representation and advanced SOM visualization, in: Proc. 11th Int'l Conf. on Discovery Science (DS-2008), Vol. LNCS 5255/2008 of Lecture Notes in Computer Science, Springer, Budapest, Hungary, 2008, pp. 259–271.

[14] E. Merényi, J. Taylor, A. Isella, Deep data: discovery and visualization application to hyperspectral ALMA imagery, Proc. Int. Astron. Union 12 (S325) (2016) 281–290, https://doi.org/10.1017/S1743921317000175.

[15] P. O'Driscoll, E. Merényi, R. Grossman, Using spatial characteristics to aid automation of SOM segmentation of functional image data, 2017 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, 2017, pp. 1–8, https://doi.org/10.1109/WSOM.2017.8020012.

[16] J.A. Lee, D.H. Peluffo-Ordóñez, M. Verleysen, Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure, Neurocomputing 169 (2015) 246–261, learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis. doi:10.1016/j.neucom.2014.12.095.

[17] J.N. Böhm, P. Berens, D. Kobak, A unifying perspective on neighbor embeddings along the attraction-repulsion spectrum, arXiv preprint arXiv:2007.08902 (2020).

[18] J. Lachmair, E. Merényi, M. Porrmann, U. Rückert, A reconfigurable neuroprocessor for self-organizing feature maps, Neurocomputing 112 (2013) 189–199, https://doi.org/10.1016/j.neucom.2012.11.045.

[19] L. Zhang, E. Merényi, Weighted differential topographic function: A refinement of the topographic function, in: Proc. 14th European Symposium on Artificial Neural Networks (ESANN'2006), 2006, pp. 13–18.

[20] T. Villmann, R. Der, M. Herrmann, T. Martinetz, Topology preservation in self-organizing feature maps: exact definition and measurement, Neural Networks IEEE Trans. 8 (2) (1997) 256–266, https://doi.org/10.1109/72.557663.

[21] T. Martinetz, K. Schulten, Topology representing networks, Neural Networks 7 (3) (1994) 507–522.

[22] B.N. Delaunay, Sur la Sphère Vide, Bull. Acad. Sci. USSR (1934) 793–800.

[23] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recogn. 46 (1) (2013) 243–256.

[24] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[25] J. Bezdek, N. Pal, Some new indexes of cluster validity, IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) 28 (3) (1998) 301–315, https://doi.org/10.1109/3477.678624.

[26] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979) 224–227. doi:10.1109/TPAMI.1979.4766909.

[27] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218, https://doi.org/10.1007/BF01908075.

[28] P. Jaccard, The distribution of the flora in the alpine zone, New Phytol. 11 (2) (1912) 37–50, https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[29] L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech: Theory Exp. 2005 (09) (2005) P09008.

[30] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.

[31] G.C. Linderman, S. Steinerberger, Clustering with t-sne, provably, SIAM J. Math. Data Sci. 1 (2) (2019) 313–332.

[32] S.A. Nene, S.K. Nayar, H. Murase, Columbia object image library (coil-20) (Tech. rep.), Columbia University, 1996.

[33] Y. LeCun, C. Cortes, MNIST handwritten digit database (2010). URL: http://yann.lecun.com/exdb/mnist/.

[34] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017). arXiv:cs.LG/1708.07747.

[35] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, D. Ha, Deep learning for classical japanese literature (2018). arXiv:cs.CV/1812.01718.

[36] A.C. Belkina, C.O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, J.E. Snyder-Cappione, Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets, Nat. Commun. 10 (1) (2019) 1–12.

[37] E. Merényi, K. Taşdemir, L. Zhang, Learning highly structured manifolds: harnessing the power of SOMs, in: M. Biehl, B. Hammer, M. Verleysen, T. Villmann (Eds.), Similarity based clustering, Lecture Notes in Computer Science, LNAI 5400, Springer-Verlag, 2009, pp. 138–168.

[38] J. Cohen, Statistical power analysis for the behavioral sciences, Routledge, 2013.

[39] L. Van Der Maaten, Accelerating t-sne using tree-based algorithms, J. Mach. Learn. Res. 15 (1) (2014) 3221–3245.

[40] G.C. Linderman, M. Rachh, J.G. Hoskins, S. Steinerberger, Y. Kluger, Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data, Nat. Methods 16 (3) (2019) 243–245.

[41] D. Kobak, G.C. Linderman, Initialization is critical for preserving global data structure in both t-sne and umap, Nat. Biotechnol. 39 (2) (2021) 156–157.

[42] T. Hastie, R. Tibshirani, Generalized Additive Models, Stat. Sci. 1 (3) (1986) 297–310, https://doi.org/10.1214/ss/1177013604.

[43] S. Li, H. Lin, Z. Zang, L. Wu, J. Xia, S.Z. Li, Invertible manifold learning for dimension reduction, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 713–728.

Josh Taylor is a Postdoctoral Fellow at The University of Texas at Austin. His research focuses on methodological development for scientific machine learning (particularly for astronomy and astrophysics), including: data-driven parameterization and assessment of unsupervised learning tasks, clustering with neural maps, and high-dimensional non-parametric statistical inference. He received the B.A. (Mathematics), M.A. and Ph.D. (Statistics) degrees from Rice University, Houston, Texas, where he also served as a Postdoctoral Researcher and Lecturer of Statistics.
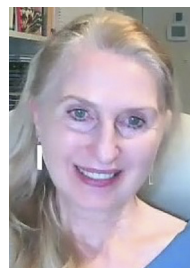
Erzsébet Merényi is a research professor in the Departments of Statistics, and Electrical and Computer Engineering, Rice University, Houston, Texas. Her current interests focus on artificial neural networks, self-organized learning of manifold structure, pattern discovery and classification in high-dimensional data, data fusion, data mining, and application to information extraction from multi- and hyperspectral remote sensing imagery of Earth, other planets and astrophysical objects, hyperspectral imagery of biological tissues, medical imagery, and multi-variate medical data. She received an M.Sc. in mathematics and Ph.D. in computational science at Szeged (Attila József) University, Hungary.