ChatGPT

Evaluation of Dimension-Reduction Methods (Huang et al. 2022)

Huang *et al.* propose five complementary strategies to evaluate DR methods on transcriptomic data [1] [2]. Each strategy has a clear purpose, specific implementation steps, and uses one or more quantitative metrics or illustrative datasets. Below we summarize each strategy, how it was carried out, and what was learned.

Local structure preservation

Local structure preservation checks whether nearby points or same-class points remain neighbors after embedding [3]. The authors implement this in two ways:

- **Supervised (labeled)**. Using labeled data (x_i, y_i), they project the high-dimensional points x_i into 2D with each DR method. Then they train a classifier (typically SVM with an RBF kernel or k-NN with $k = 5$ [4]) on a subset of the low-d embedding and test it on held-out points. High classification accuracy means points of the same label stayed close in 2D (homophily) [5]. For example, on the MNIST digit dataset, t-SNE, UMAP, TriMap and PaCMAP all gave high SVM/k-NN accuracy (closer to the original cluster separability), whereas ForceAtlas2 performed poorly [6]. Across many single-cell datasets, nearly all DR methods (except ForceAtlas2) achieved high accuracy: t-SNE, UMAP, PaCMAP and related methods typically scored best [7].
- **Unsupervised (no labels)**. When ground-truth labels are not available, they measure how well local neighborhoods are preserved. Concretely, for each point i, they find its $k = 5$ nearest neighbors in high-dimensional space $N(i)$ and its $k = 5$ neighbors in the 2D embedding $N'(i)$. They compute the fraction $|N(i) \cap N'(i)|/k$ for each i and average over all points [8]. A higher average means more of each point's original neighbors remain neighbors. Using this metric, t-SNE and its accelerated version (art-SNE) preserved the largest fraction of neighbors, while PCA (a purely linear method) preserved the fewest [9].

Key findings: Methods designed for local clustering (t-SNE, UMAP, PaCMAP) excel at keeping class-neighbors together, yielding high classification accuracy and neighbor preservation [7] [9]. In contrast, ForceAtlas2 and PHATE underperform on these local metrics.

Global structure preservation

Global structure preservation checks whether overall relationships and cluster arrangements are maintained. The authors use both qualitative and quantitative evaluations for this:

- **Qualitative examples.** They test on datasets with known large-scale structure (e.g. a synthetic “mammoth” point cloud, and simulated Gaussian-cluster data with a linear or hierarchical arrangement). By projecting these into 2D, they visually inspect whether connected parts stay together or false separations appear [10]. For instance, on the 3D “Mammoth” dataset, t-SNE and

UMAP fragmented the shape into disconnected clusters, whereas methods like TriMap and ForceAtlas2 preserved the overall form [10]. Similarly, on a Gaussian-linear test, methods like TriMap, PaCMAP and ForceAtlas2 showed a smooth color gradient reflecting the original high-dim structure, while t-SNE/UMAP did not [11]. On a harder hierarchical Gaussian dataset, ForceAtlas2 (and PaCMAP) did best at keeping micro-clusters of the same meso-cluster nearby [12].

Figure: Qualitative examples of global-structure evaluation. In Huang *et al.*, a 3D “mammoth” shape (a) is projected into 2D by different DR methods (b); TriMap and ForceAtlas2 preserve the mammoth’s connected outline, whereas t-SNE and UMAP break it into disjoint pieces [10]. Similar tests on synthetic Gaussian data (c,d) highlight which methods retain large-scale cluster arrangements [11].

- **Quantitative metrics.** The paper defines several metrics that compare distances or cluster relationships in high- and low-dimensional spaces:
- **Random Triplet Accuracy:** Sample many random triplets of points and check if the ordering of their pairwise distances is the same in high-D and low-D (i.e. if the closest pair remains closest, etc). The metric is the fraction of triplets with preserved order [13]. (They use a sample of triplets for tractability.) A higher triplet accuracy means global distance relations are well-preserved. On this metric, TriMap scored highest (consistent with its optimizing a triplet loss), while PCA, PaCMAP and ForceAtlas2 also performed well. In contrast, t-SNE, UMAP and PHATE scored relatively lower [14].
- **Distance Spearman Correlation:** Compute all (or many) pairwise distances in the original data and in the embedding, then compute Spearman’s rank correlation between these two distance vectors [15]. A high correlation means the global distance ranking is maintained. Again, TriMap, PCA, PaCMAP and ForceAtlas2 scored well on most datasets, while t-SNE, UMAP (and art-SNE) tended to score worse [15].
- **Cluster-centric metrics:** They also measure how well inter-cluster relationships are preserved by comparing cluster centroids. One metric is **k-Nearest Class Preservation**: for each class centroid, see how many of its k nearest-neighbor *centroids* in high-D remain its nearest neighbors in 2D [16]. (They set $k = \lfloor (C+2)/4 \rfloor$, where C is the number of classes, to adapt to dataset size.) Another is **Centroid Distance Correlation**: compute the Spearman correlation between the set of all inter-centroid distances in high-D and in low-D [17]. These metrics again favored methods like TriMap and ForceAtlas2. On k-nearest class preservation, PCA and ForceAtlas2 were most consistent across datasets [18], and on centroid-distance correlation TriMap and ForceAtlas2 scored highest [17].

Key findings: Methods optimized for global geometry (TriMap, ForceAtlas2, PaCMAP) outperform local-only methods in these global metrics [14] [17]. By contrast, t-SNE and UMAP – which focus on local neighborhoods – often produce high-quality local clusters but badly distort the large-scale layout (creating “false” clusters) [10] [15].

Sensitivity to parameter choices

This evaluation checks whether small changes in DR tuning parameters lead to large changes in the embedding (which would indicate instability) [2]. The steps are:

1. **Select datasets with known structure** (e.g. Kazer *et al.* [immune cells] and Stuart *et al.* [hematopoietic progenitors]) [19].

2. **Vary key parameters** for each algorithm. For example, for t-SNE they varied *perplexity*, for UMAP varied *n_neighbors*, and for TriMap varied *n_inliers*. Each parameter controls the “spread” or locality of the embedding forces [19].
3. **Generate multiple embeddings.** For each choice of parameter (within a reasonable range), run the DR algorithm from the same preprocessed input (they used PCA to 70 PCs initially).
4. **Compare cluster relationships.** Examine how distances between known cell-type clusters change. In their figures, changing t-SNE’s perplexity or UMAP’s n_neighbors caused the distance between certain cell clusters (e.g. two types of dendritic cells, or two progenitor types) to vary dramatically [19]. In some cases a single cell type even split into two clusters under one setting but not another (black circles in Fig.5a) [19].
5. **Interpret sensitivity.** Large shifts in cluster arrangement indicate that the DR method is sensitive to its parameters. The authors note that t-SNE and UMAP showed strong sensitivity (cluster distances moved a lot) [19] [20]. By contrast, TriMap and PaCMAP embeddings were more stable across parameter changes in these examples (their colored clusters stayed in roughly the same relative positions) [20].

Figure: Example of parameter-sensitivity tests (from Huang *et al.*). Changing t-SNE’s perplexity or UMAP’s *n_neighbors* dramatically alters the distance between certain cell clusters (circled), demonstrating instability [19]. In the lower panel, UMAP’s output on two progenitor cell types shifts markedly as *n_neighbors* changes, whereas TriMap’s output is more consistent [20].

Key point: The authors do not compute a single metric here, but emphasize visual inspection. Methods whose embeddings change dramatically with parameter tweaks are deemed “sensitive”. Their results showed that popular methods like t-SNE and UMAP are often highly sensitive, whereas TriMap/PaCMAP tend to be more robust to parameter variation [19] [20].

Sensitivity to preprocessing choices

This test examines how the embedding changes when the input preprocessing is varied. Transcriptomic data usually undergo log-normalization and PCA before DR, but alternative schemes exist. The study’s steps were:

1. **Vary PCA dimensionality.** After log-normalizing the data, they applied PCA to reduce to different numbers of principal components (e.g. 30, 50, 70, 100) before running the DR methods [21]. They tested this on two scRNA-seq datasets (Kazer and Stuart). For each method and PC setting, they generated the 2D embedding.
2. **Compare cluster arrangement.** They focused on known related cell types (e.g. two dendritic cell subtypes) and measured the distance between those clusters in the embedding. As shown in Fig.6, t-SNE and UMAP produced very different cluster distances when using 50 vs 70 PCs, whereas TriMap and PaCMAP embeddings were nearly unchanged [21]. In one case, UMAP with 50 PCs placed two related progenitor clusters far apart, whereas with 70 PCs they were close (splitting one cluster into two) [21].
3. **Vary preprocessing pipeline.** Beyond PCA, they tested three pipelines on one dataset: (a) **raw data + PCA**, (b) **log-normalization + PCA**, and (c) **GLM-PCA** (a specialized PCA for count data) [22]. Again, they applied each DR method to each preprocessed input.
4. **Check stability of results.** In Fig.7, t-SNE and UMAP embeddings changed dramatically between the three pipelines – the distance between two dendritic-cell clusters (mDC vs pDC) varied wildly and

many outliers appeared under GLM-PCA [22] [23]. PaCMAP's embeddings, however, showed only minor differences across preps.

Figure: Example of preprocessing-sensitivity tests. On the Stuart *et al.* dataset, three preprocessing pipelines were tried (raw+PCA, log-normalized+PCA, GLM-PCA). The black-circled DC clusters (mDC vs pDC) move relative to each other in t-SNE/UMAP embeddings as the pipeline changes, whereas PaCMAP's result is more consistent [22].

Key point: An ideal DR method would be robust to such choices. Huang *et al.* found that t-SNE and UMAP were *not* robust: their embeddings (cluster distances, outlier count, etc.) changed a lot with different preprocessing [22]. By contrast, TriMap and especially PaCMAP were relatively insensitive to the number of PCs and to using GLM-PCA [21] [22].

Computational efficiency and scalability

This strategy measures running time of each method on datasets of increasing size [24]. The procedure was:

1. **Select datasets of various sizes.** They used multiple scRNA-seq datasets (from a few thousand to >1 million cells) discussed throughout the paper [24].
2. **Time each DR algorithm.** For each method, run it with default settings to embed each dataset from $N \times d$ to $N \times 2$, recording the elapsed time. (For methods like ForceAtlas2 that require a graph, they included graph construction time for fairness [24] [25].)
3. **Compare run times.** In Fig.8, PaCMAP consistently took the least time on almost all datasets, especially the largest ones [26]. UMAP, TriMap and t-SNE also scaled reasonably well up to ~10^5–10^6 cells (t-SNE using a fast implementation) [26]. In contrast, art-SNE and PHATE were much slower: art-SNE ran out of memory on the ~10^6-cell datasets, and PHATE failed to finish within 24 hours on those large sets [26] [25].

Figure: Running time vs. sample size for various DR algorithms (log–log scale). PaCMAP (purple) is fastest across all sizes, UMAP (green) and TriMap (red) also scale well. art-SNE (brown) and PHATE (pink) become impractical on the largest datasets [26].

Key findings: PaCMAP is the fastest for large datasets, enabling rapid exploration of very large single-cell datasets [26]. UMAP, TriMap and optimized t-SNE are also reasonably fast. ForceAtlas2, art-SNE and PHATE are significantly slower and may even fail on multi-million-cell data under typical resource limits [26] [25]. This has practical importance: a faster method allows users to rerun DR many times (e.g. for parameter tuning) and handle modern large-scale transcriptomic datasets.

Summary: Huang *et al.*'s five-fold evaluation framework systematically probes different aspects of DR quality. Local-preservation tests (supervised kNN/SVM accuracy and neighbor overlap) identify how well nearby points stay together [6] [9]. Global-preservation tests (triplet/distance correlations and cluster-centroid metrics) quantify how well overall geometry is maintained [13] [17]. Sensitivity analyses reveal which algorithms yield stable embeddings across parameter or preprocessing changes [19] [22]. And runtime comparisons highlight practical efficiency and scalability [24]. Together, these evaluations show that methods like TriMap, ForceAtlas2 and PaCMAP are strong on global structure and robustness, while t-SNE/UMAP excel at local clustering but often distort global layouts and require careful parameter tuning [10] [19].

Sources: All information above is drawn from the sections “Evaluation 1–5” of Huang *et al.* (2022), *Communications Biology*, which details the purpose, methodology, metrics, and findings for each evaluation strategy [3] [24].

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization | Communications Biology
https://www.nature.com/articles/s42003-022-03628-x?error=cookies_not_supported&code=c3c4d1d6-c5e6-467f-b6f0-0158c2862ed0