

Do Better Embeddings Lead to Better Judgments? A User Study of PCA and UMAP for Single-Cell Data

Helena Qing Huang ^{*}, Kris Sankaran [†]

ABSTRACT

Dimensionality-reduction visualizations are essential to modern biology, but the same data can yield different embeddings depending on algorithm and hyperparameter choices. Whether these choices influence analyst conclusions, and whether more faithful embeddings yield better conclusions, is unknown. We ran an exploratory within-subject study with 12 participants comparing PCA, UMAP set to default hyperparameters, and a tuned version of UMAP across clustering and trajectory tasks grounded in single-cell analysis workflows. Both UMAP conditions produced higher accuracy and confidence than PCA and were strongly preferred, and completion time did not differ significantly. However, tuning did not improve accuracy over default UMAP and higher confidence did not correspond to higher accuracy. These findings suggest that computational embedding fidelity and perceptual usefulness are related but distinct.

Keywords: dimensionality reduction, genomics, user study

1 INTRODUCTION

Dimensionality reduction (DR) visualizations project high-dimensional data into lower-dimensional views. They are central to genomics research, where they help navigate molecular profiles [1], and are built into widely used software. Alternative projections emphasize different aspects of the original structure [7]. Principal Component Analysis (PCA), for instance, is a linear projection that captures dominant axes of variance, while nonlinear methods such as Uniform Manifold Approximation and Projection (UMAP) prioritize local neighborhood preservation [4]. Although these embeddings are often interpreted as spatial maps of cellular similarity, apparent features such as cluster separation, inter-cluster distance, compactness, and global arrangement may not faithfully correspond to relationships in the gene-expression space.

Recent work has proposed alternatives that better reflect distances in the original space. For example, scDEED tunes hyperparameters to improve neighborhood distance preservation [6]. Prior perceptual studies have shown that DR methods can influence users' accuracy, completion time, and interpretation of cluster structure [2, 3] and have identified systematic differences in how people perceive cluster boundaries, density, similarity, and outliers across projection methods. However, they have largely examined generic high-dimensional datasets and DR algorithms. It therefore remains unclear whether improving embedding fidelity in biologically-grounded DR visualization also improves perceptual judgments.

In single-cell workflows, PCA scores provide the pre-embedding representation from which nonlinear embeddings such as UMAP are constructed. We therefore use PCA as a linear reference for examining what the nonlinear embedding changes perceptually, while Default versus Tuned UMAP isolates whether improving the reliability of that embedding further improves judgments. We investigate this question through an exploratory within-subject study compar-

ing PCA, UMAP, and an scDEED-tuned UMAP across single-cell interpretation tasks:

- **RQ1:** How does perception accuracy vary across DR encodings in a single-cell workflow?
- **RQ2:** How do encodings affect confidence, completion time, and self-reported preference?
- **RQ3:** Does improved distance reliability translate to higher perceptual accuracy?

2 METHODOLOGY

Tasks and stimuli. We focus on two visual judgments that analysts make when reading DR embeddings of single cell data [1]. Phase A (T1–T4) targets clustering judgments about discrete population structure using cord blood cells profiled by CITE-seq, chosen for their well-characterized cell-type labels. The four tasks asked participants to count populations (T1), name the most abundant type (T2), name the type with highest within-type similarity (T3), and mark regions with hypothesized distinct genetic profiles (T4). Phase B (T5, T6) targets trajectory judgments – reasoning about how well cell populations differentiate – using an HSPC dataset, chosen for its well-mapped differentiation pathways. The two tasks asked participants to order LTHSC, MPP, and LMPP developmentally (T5) and to name the ancestors of CMP (T6). All visualizations (Fig. 1) were generated in R. The data generation process, full task descriptions, and prompts are provided in the supplemental material as Table 1.

Procedure and analysis. Every task is followed by a 5-point Likert confidence rating (1 = not at all confident, 5 = very confident). Reading confidence alongside accuracy lets us detect *miscalibration*. After both phases, participants completed a post-study survey on their overall preferred visualization and open-ended feedback. Twelve participants completed a 75-minute within-subject study; all self-reported prior course experience with biology and reading scientific graphics. After a visualization tutorial, participants began T1–T6 under counterbalanced visualization conditions, rated their confidence after each response, and completed a post-study interview. All participants provided informed consent.

We scored T1–T3, T5, and T6 against dataset-derived ground truth. T4 is open-ended without ground truth. Accuracy was analyzed with mixed-effects logistic regression, confidence with an ordinal cumulative-link mixed model, and log-transformed completion time with a linear mixed-effects model. Models included the visualization condition and task as fixed effects and participant as a random effect. We read these results cautiously because of the small sample size and task-by-condition cells near the ceiling or floor.

3 RESULTS AND DISCUSSION

RQ1: Accuracy improved with UMAP, but not uniformly. Table 1 reports model-based contrasts and Fig. 2 shows the task-level distributions. Both UMAP conditions raised the odds of an exact-match response compared to PCA. The gap was largest in developmental ordering (T5), for which 5/12 were correct for PCA, 12/12 for Default UMAP, and 11/12 for Tuned UMAP. UMAP also helped on population abundance (T2) and lineage ancestry (T6), though only modestly for T6. This advantage was uneven. Population count was similar across conditions (8/12, 7/12, 8/12), and for within-type similarity (T3), Tuned UMAP scored worst (1/12, versus 5/12 for

^{*}e-mail: qhuang239@wisc.edu

[†]e-mail: ksankaran@wisc.edu

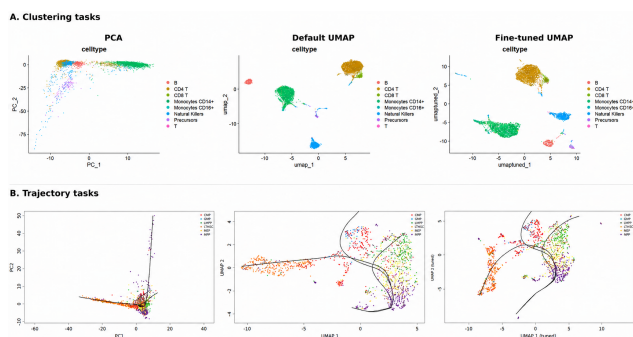


Figure 1: PCA, Default UMAP, and Fine-tuned UMAP stimuli for (A) clustering and (B) trajectory tasks.

Table 1: Model-based contrasts. Accuracy and confidence are given as odds ratios (OR), while completion time is given as a time ratio. Models control for task and include participant random effects.

Outcome	Contrast	Effect [95% CI]	<i>p</i>
Accuracy	Default UMAP vs. PCA	OR = 3.67 [1.52, 8.90]	0.004
	Tuned UMAP vs. PCA	OR = 3.08 [1.28, 7.43]	0.012
	Tuned vs. Default UMAP	OR = 0.82 [0.35, 1.95]	0.700
Confidence	Default UMAP vs. PCA	OR = 3.10 [1.57, 6.11]	0.001
	Tuned UMAP vs. PCA	OR = 2.34 [1.20, 4.57]	0.013
	Tuned vs. Default UMAP	OR = 0.76 [0.39, 1.45]	0.400
Time	Default UMAP vs. PCA	Ratio = 0.91 [0.76, 1.08]	0.300
	Tuned UMAP vs. PCA	Ratio = 1.11 [0.93, 1.33]	0.200
	Tuned vs. Default UMAP	Ratio = 1.23 [1.03, 1.47]	0.023

PCA and 6/12 for Default UMAP). The significant condition contrasts are therefore average effects across the five tasks, not evidence that UMAP improved every judgment. Because each task used few stimuli, we interpret the task-level differences descriptively. T4 had no ground truth and was excluded from the scored accuracy. Marker placement was most dispersed under PCA, where absent boundaries forced participants to impose their own interpretive frames, and more convergent under both UMAP conditions, where visible islands anchored responses (corresponding figure refers to supplement).

RQ2: UMAP elicited higher confidence and preference, with no detected time difference. Both UMAP conditions raised the odds of a higher confidence rating relative to PCA. Completion time showed no statistically significant difference. Preference results were similar. For clustering tasks, all 12 participants preferred a UMAP condition: 9 the Default UMAP and 3 Tuned. For trajectory tasks, 10 of 12 preferred UMAP, divided evenly between the two types. We did not formally test confidence calibration, and several task-level cells paired high confidence with low exact-match accuracy. These results align with prior results on generic datasets [5].

RQ3: Fine-tuning did not improve task accuracy. We found no evidence that tuning improved exact-match accuracy over Default UMAP. The wide confidence interval does not demonstrate equivalence; it is compatible with both a disadvantage and a moderate advantage for the hyperparameter-tuned condition. Within the stimuli examined here, optimizing UMAP for the scDEED reliability criterion did not translate into better task performance. Embedding fidelity and perceptual utility may therefore capture related but distinct properties of a DR visualization. Participants may base their judgments on visible separation, density, compactness, or trajectory shape, and tuning can change these cues in ways that do not improve every task. Further studies with additional datasets, task instances, and tuning strategies are needed before generalizing this result.



Figure 2: Perceptual outcomes by task and visualization condition. (A) Exact-match accuracy; error bars are 95% Wilson CIs. (B) Completion time in seconds. (C) Confidence ratings (1–5). (D) Post-study visualization preference by task phase. T4 is excluded from (A) and (C) as it had no ground truth.

4 LIMITATION AND FUTURE WORK

These findings are exploratory. With 12 participants and several task-condition cells at or near floor or ceiling, the precision of the model estimates is limited. Each judgment type was represented by relatively few stimuli across two datasets, so descriptive task-level differences may reflect those datasets or prompts rather than the task categories. In addition, exact-match scoring for T5 and T6 discarded potentially informative partial-credit differences, and we evaluated only one reliability-tuning method, limiting the generality of the Default vs. Tuned UMAP comparison. Future work should recruit a larger and more domain-experienced sample, replicate each judgment across multiple datasets and task instances, analyze partial-credit outcomes alongside exact-match accuracy, and test visualization-by-task interactions and confidence calibration directly.

REFERENCES

- [1] R. A. Amezquita, A. T. Lun, E. Becht, V. J. Carey, L. N. Carpp, L. Geistlinger et al. Orchestrating single-cell analysis with bioconductor. *Nature methods*, 17(2):137–145, 2020. doi: [10.1038/s41592-019-0654-x](https://doi.org/10.1038/s41592-019-0654-x) 1
- [2] S. Doh, H. Jeon, S. Shin, G. J. Quadri, N. W. Kim, and J. Seo. Understanding bias in perceiving dimensionality reduction projections. *arXiv preprint arXiv:2507.20805*, 2025. doi: [10.48550/arXiv.2507.20805](https://doi.org/10.48550/arXiv.2507.20805) 1
- [3] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):81–94, 2015. doi: [10.1109/TVCG.2014.2330617](https://doi.org/10.1109/TVCG.2014.2330617) 1
- [4] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: [10.48550/arXiv.1802.03426](https://doi.org/10.48550/arXiv.1802.03426) 1
- [5] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):529–539, Jan. 2022. doi: [10.1109/TVCG.2021.3114694](https://doi.org/10.1109/TVCG.2021.3114694) 2
- [6] L. Xia, C. Lee, and J. J. Li. Statistical method sceed for detecting dubious 2d single-cell embeddings and optimizing t-sne and umap hyperparameters. *Nature Communications*, 15(1):1753, 2024. doi: [10.1038/s41467-024-45891-y](https://doi.org/10.1038/s41467-024-45891-y) 1
- [7] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, and X. Chen. A comparison for dimensionality reduction methods of single-cell rna-seq data. *Frontiers in Genetics*, Volume 12 - 2021, 2021. doi: [10.3389/fgene.2021.646936](https://doi.org/10.3389/fgene.2021.646936) 1