Overcoming Analysis Paralysis: Comparing Dimension Reduction algorithms across Local and Global Data Structure

Jamie S. McCarthy[1], Geoffrey Kraker[1], JF Tirados[1], Andrea Valle[1], Chris Ciccolella[1].
1: OMIQ, Dotmatics Ltd.

Dotmatics OMIQ
www.omiq.ai

Introduction

- Dimension reduction (DR) is widely used in high-dimensional cytometry analysis; **UMAP** and **opt-SNE** are the most common choices in OMIQ.
- Selecting the right tool is challenging: each algorithm has settings that affect its output, and the right choice depends on your analysis question (local vs global phenotyping or continuous gradients) and dataset size.
- **Local structure** captures relationships between nearby cells (e.g. CCR7 expression on CD4+ T cells); **Global structure** preserves the overall geometry and relative positions of populations across the full dataset (e.g. the distinction between B cells and monocytes).
- Here we compared six DR algorithms in OMIQ, exploring how each performs across local vs global structure, hyperparameter (settings) changes, input dimensionality, and scalability.

Methods

- To test DR we used healthy control PBMCs from a spectral flow cytometry experiment: 4 donor samples with 40 parameters available for inclusion.
- Data were arcsinh scaled, cleaned with PeacoQC followed by manual gating, and clustered with FlowSOM (10 or 30 clusters) for overlay onto DR embeddings.
- Embeddings were assessed visually using overlay plots (cluster identity) and coloured continuous plots (marker expression, e.g. CD4/CD8).

Publicly available data used from Flow Repository (ID: R-FCM-Z2QV).

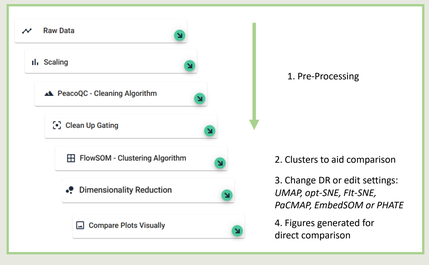


Fig 1. Data analysis methods visualised as an OMIQ workflow

Results

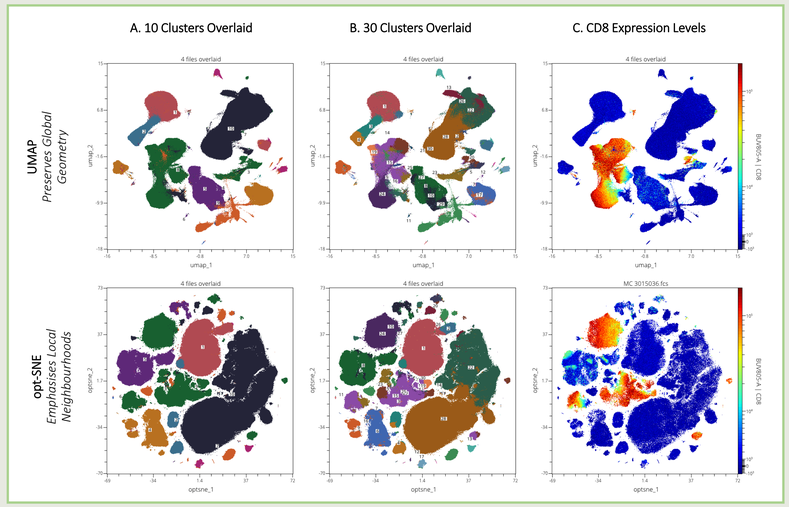


Fig 2. UMAP vs. opt-SNE: Local versus Global Structure

UMAP and opt-SNE embeddings shown three ways:
A. With 10 FlowSOM clusters for a broad overview (global-like).
B. With 30 FlowSOM clusters showing fine-grain populations (local-like).
C. CD8 expression on the Z-axis (colour); red = high. UMAP groups all CD8+ cells into one island, whilst opt-SNE splits them across many.

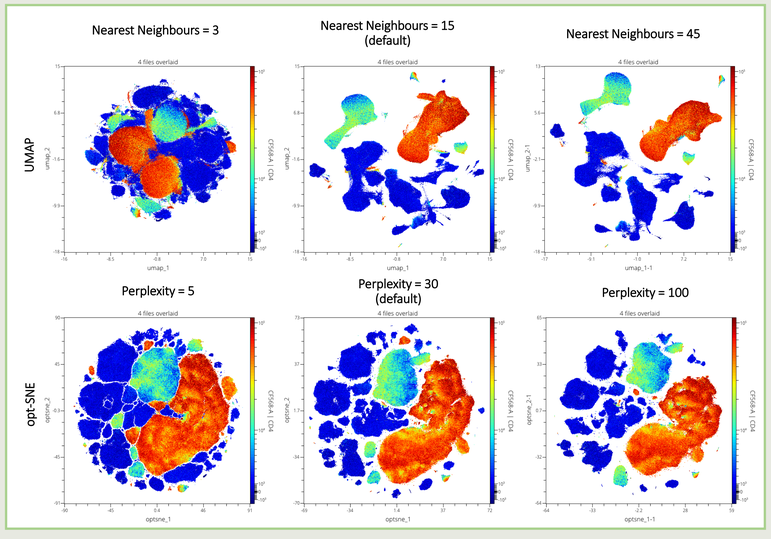


Fig 4. Algorithm Settings Impact Results

Hyperparameters shift the local/global balance within a single algorithm. UMAP's Nearest Neighbours and opt-SNE's Perplexity have similar effects: lower values are more local, higher values more global.

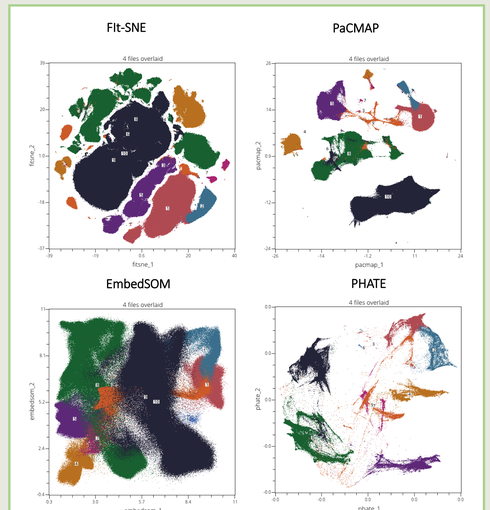


Fig 3. Other Dimension Reduction Methods

10 FlowSOM clusters overlaid across four additional algorithms.
FIt-SNE: shares origins with opt-SNE but requires more intentional setting.
PaCMAP: a blend of local and global structure.
EmbedSOM: uses self-organising maps (like FlowSOM) to capture local structure.
PHATE: best for continuous gradients/trajectories rather than phenotyping; captures global structure.

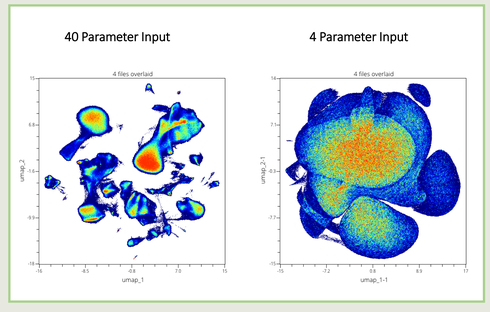


Fig 5. Input parameter count shapes embedding complexity

Two otherwise identical UMAPs were run with 40 vs 4 input parameters. More parameters increased the complexity of the embedding.

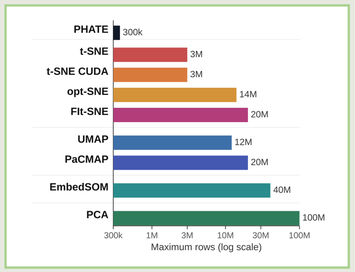


Fig 6. Algorithm scalability by input row count

DR tasks were "titrated" in OMIQ to find the row count at which each algorithm fails. Absolute thresholds are hardware-dependent, but the relative ranking reflects each algorithm's memory complexity and is preserved across compute environments.

Conclusion and Discussion

- **UMAP vs opt-SNE (Fig 2):** UMAP captures a more global view, combining subpopulations into broader islands; opt-SNE resolves local structure with distinct islands per subpopulation.
- **Settings matter (Fig 4):** UMAP's Nearest Neighbours and opt-SNE's Perplexity both shift the local/global balance within a single algorithm.
- **Input choice matters (Fig 5):** more parameters increase embedding complexity. Only include parameters relevant to your question; the rest is noise.
- **Scalability differs (Fig 6):** algorithms vary substantially in the number of rows they handle before failure. Choose based on dataset size as well as structure preference.
- **Summary (Fig 7):** algorithm selection is a tradeoff between local/global preservation and scalability.

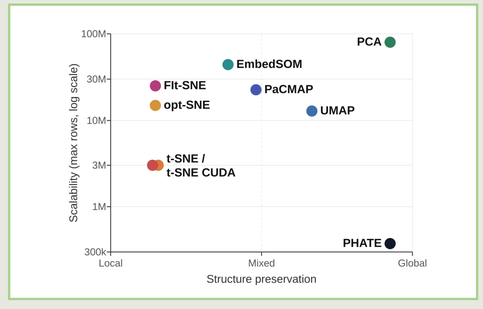


Fig 7. Algorithm overview: local/global structure preservation vs scalability (max input rows)

References

1. McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861.
2. Belkina, A.C., Ciccolella, C.O., Anno, R., Halpert, R., Spidlen, J., & Snyder-Cappione, J.E. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nature Communications, 10, 5415.
3. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., & Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods, 16, 243–245.
4. Kratochvíl, M., Koladiya, A., & Vondrášek, J. (2020). Generalized EmbedSOM on quadtree-structured self-organizing maps [version 2; peer review: 2 approved]. F1000Research, 8, 2120.
5. Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., van den Elzen, A., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. Nature Biotechnology, 37(12), 1482–1492.
6. Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. Journal of Machine Learning Research, 22(201), 1–73.

Presented at CYTO 2026