Supporting Information

SI Simulated Examples

SI.1 Trefoil Plots

For the trefoil example, the signal Y consisted of a trefoil knot embedded in three dimensions containing 500 points. $Z + \epsilon$ was constructed by adding seven superfluous dimensions and isotropic Gaussian noise. Various degrees of noise were tested ($sd = 5, 10, 15, 20, 25, 30$). The first two plots depict Trustworthiness vs. Perplexity and the trustworthiness-maximizing embeddings for the $sd = 10$ case. The third plot shows the trustworthiness-maximizing perplexity for the different degrees of noise.

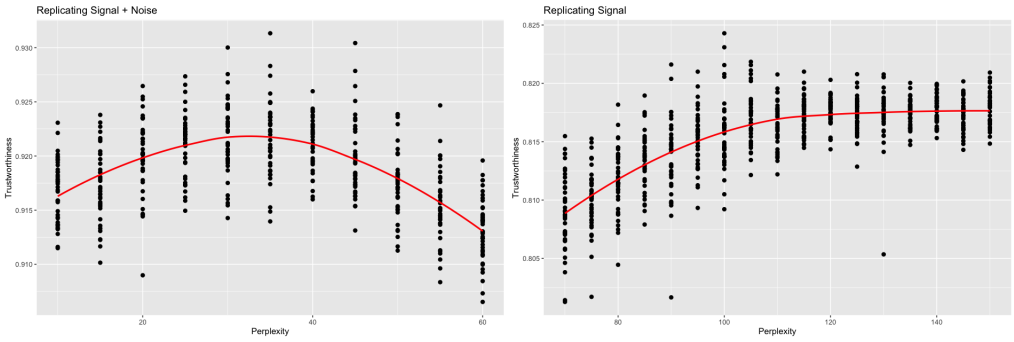

Fig S1: Trustworthiness vs. Perplexity (Trefoil)

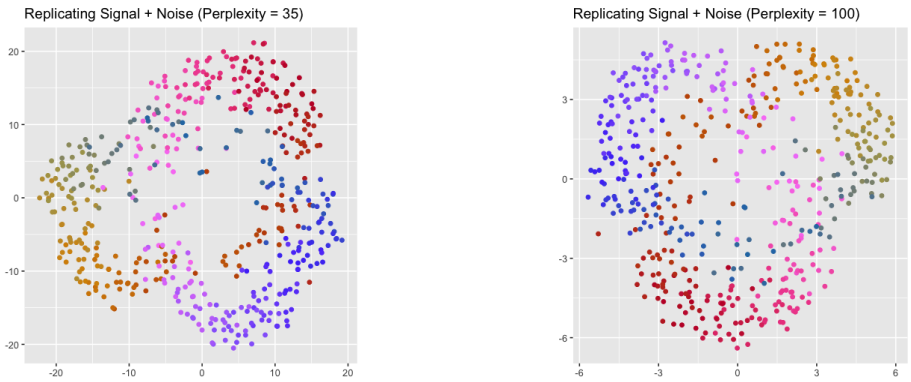

Fig S2: Trustworthiness-Maximizing Representations (Trefoil)

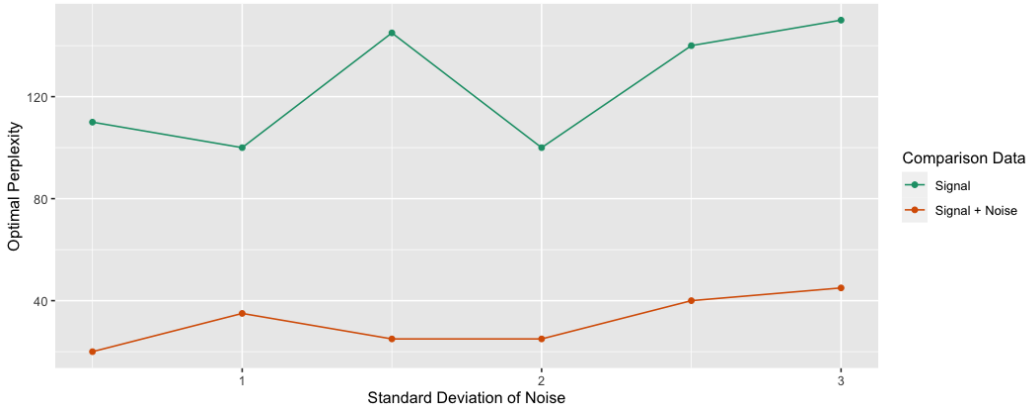

Fig S3: Optimal Perplexity (Trefoil)

SI.2 Mammoth Plots

For the mammoth example, the signal Y consisted of 500 points in three dimensions. The data was randomly sampled from the mammoth data set used in [18]. $Z + \epsilon$ was constructed by adding seven superfluous dimensions and isotropic Gaussian noise. Various degrees of noise were tested ($sd = 0.5, 1, 1.5, 2, 2.5, 3$). The first two plots depict Trustworthiness vs. Perplexity and the trustworthiness-maximizing embeddings for the $sd = 1$ case. The third plot shows the trustworthiness-maximizing perplexity for the different degrees of noise.

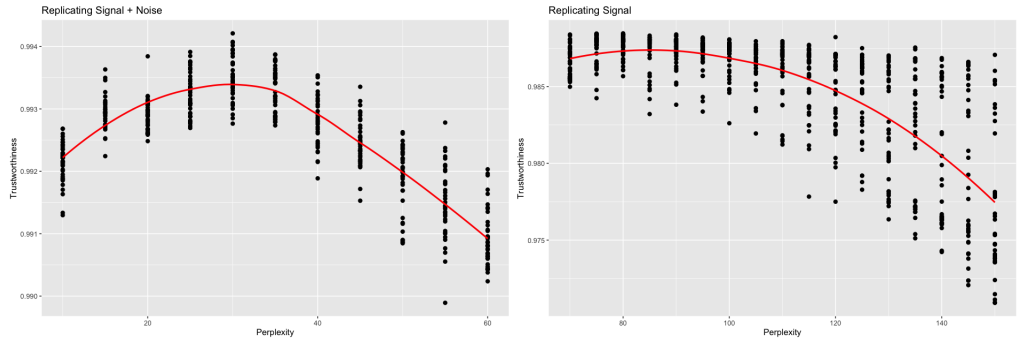

Fig S4: Trustworthiness vs. Perplexity (Mammoth)

Fig S5: Trustworthiness-Maximizing Representations (Mammoth)

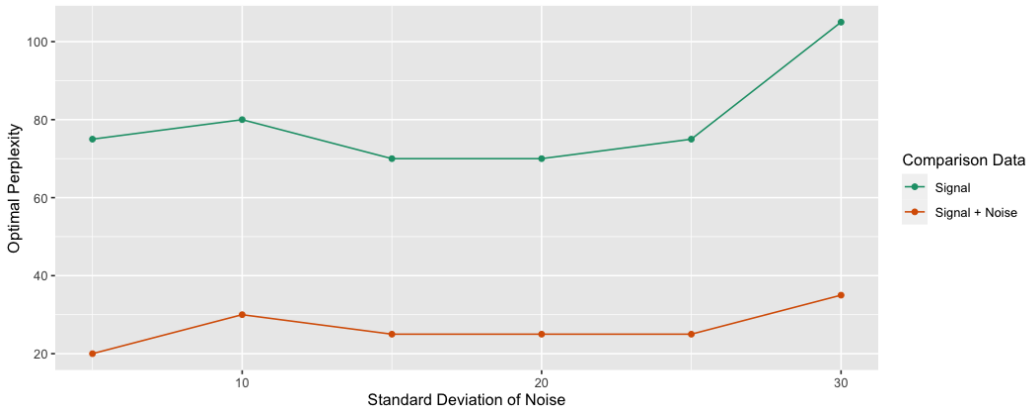

Fig S6: Optimal Perplexity (Mammoth)

SII Practical Examples

SII.1 CyTOF Data Set

The CyTOF data set contained 239,933 observations in 49 dimensions [21]. To reduce the computational load, a subset of 5,000 observations was sampled. A log transformation was followed by a PCA pre-processing step to reduce the number of dimensions to 30, which still retained 77% of the variance in the original data. The processed data set to be studied consisted of 5,000 observations in 30 dimensions. The signal was first taken to be the first five principal components, then the first eight principal components. A hierarchical clustering of the high-dimensional data was computed then projected onto the trustworthiness-maximizing representations.

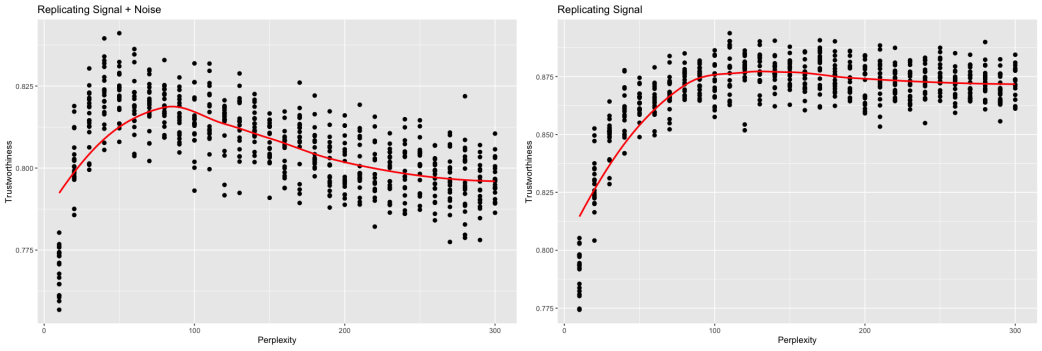

Fig S7: Trustworthiness vs. Perplexity for r = 5 (CyTOF)

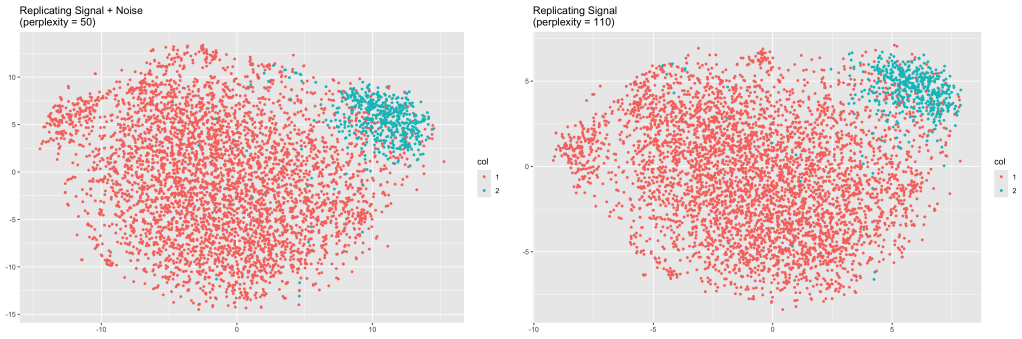

Fig S8: Trustworthiness-Maximizing Representations for r = 5 (CyTOF)

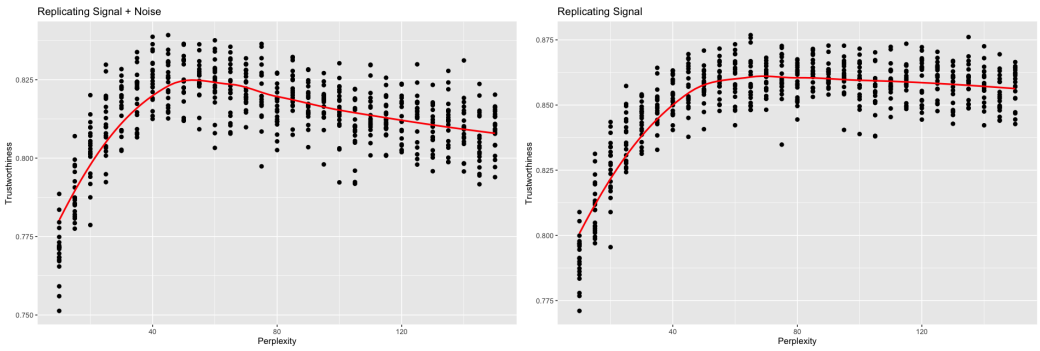

Fig S9: Trustworthiness vs. Perplexity for r = 8 (CyTOF)

Fig S10: Trustworthiness-Maximizing Representations for r = 8 (CyTOF)

SII.2 Microbiome Data Set

[22] compares the faecal microbial communities from 22 subjects using complete shotgun DNA sequencing. The original data contained 280 samples and 553 genera. To deal with a large number of near-zero readings, columns containing a large proportion of values less than 10^{-6} (60% or more) were removed. This reduced the dimension to 66. A PCA pre-processing was used to center and re-scale the data. The signal was first taken to be the first five principal components, then the first eight principal components. A k-means clustering of the high-dimensional data was computed then projected onto the trustworthiness-maximizing representations.

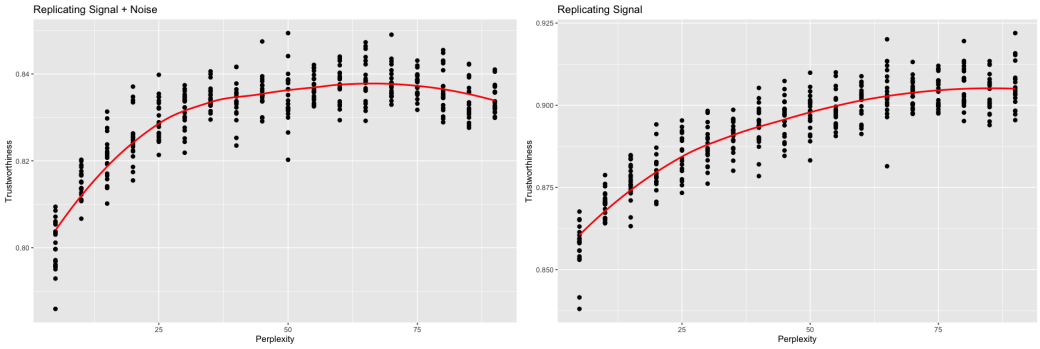

Fig S11: Trustworthiness vs. Perplexity for r = 5 (Microbiome)

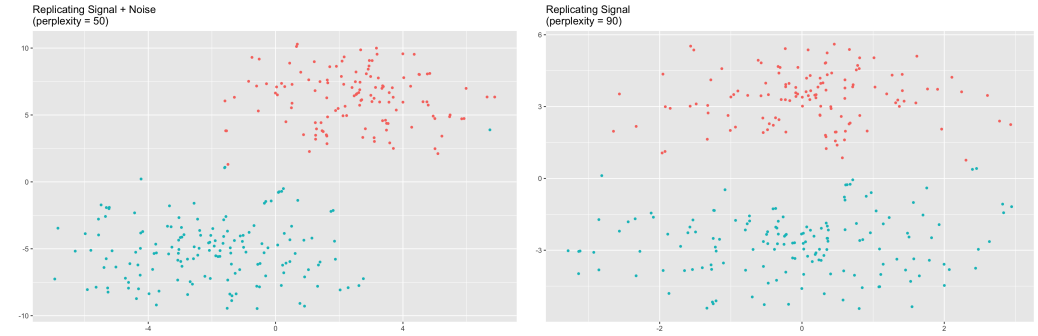

Fig S12: Trustworthiness-Maximizing Representations for r = 5 (Microbiome)

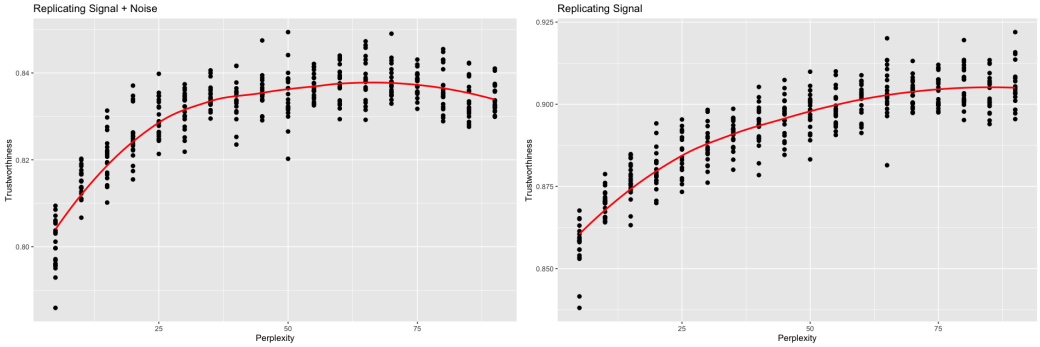

Fig S13: Trustworthiness vs. Perplexity for r = 8 (Microbiome)

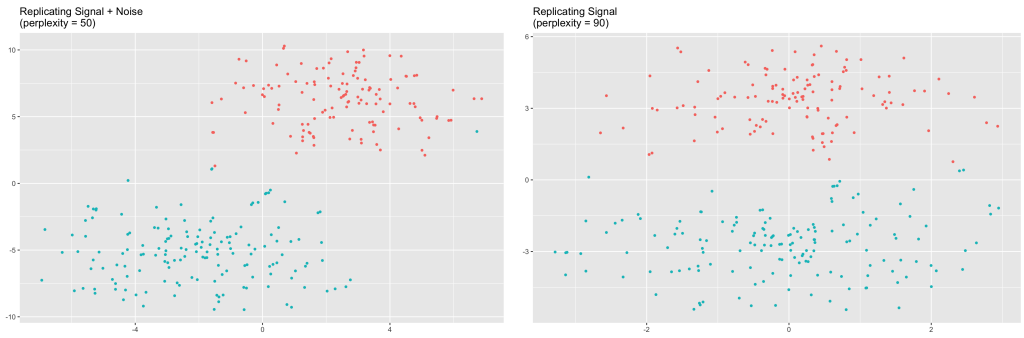

Fig S14: Trustworthiness-Maximizing Representations for r = 8 (Microbiome)

SIII PBMC Data Set

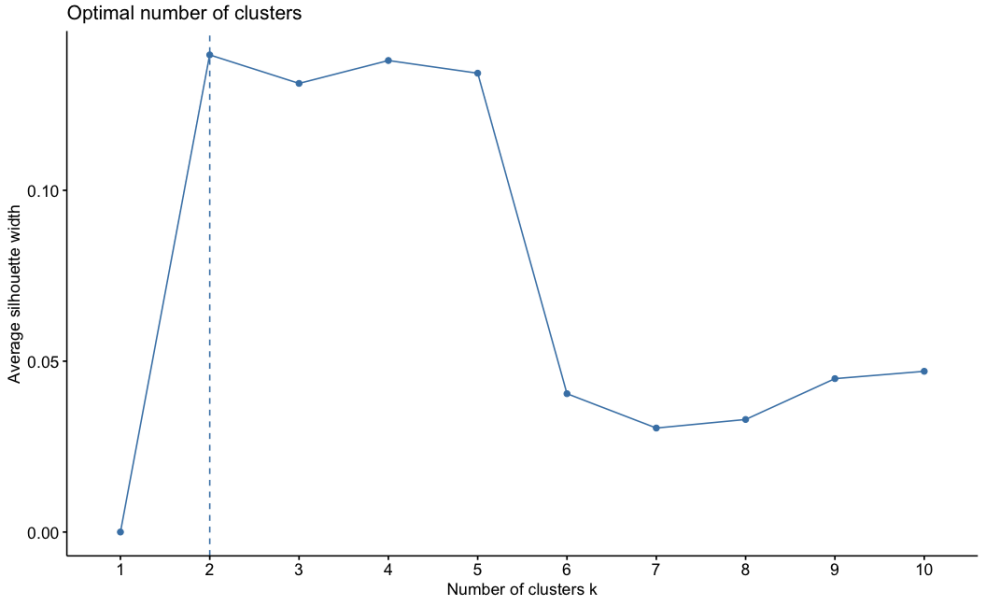

Fig S15: Average Silhouette Width for Dendritic Cells

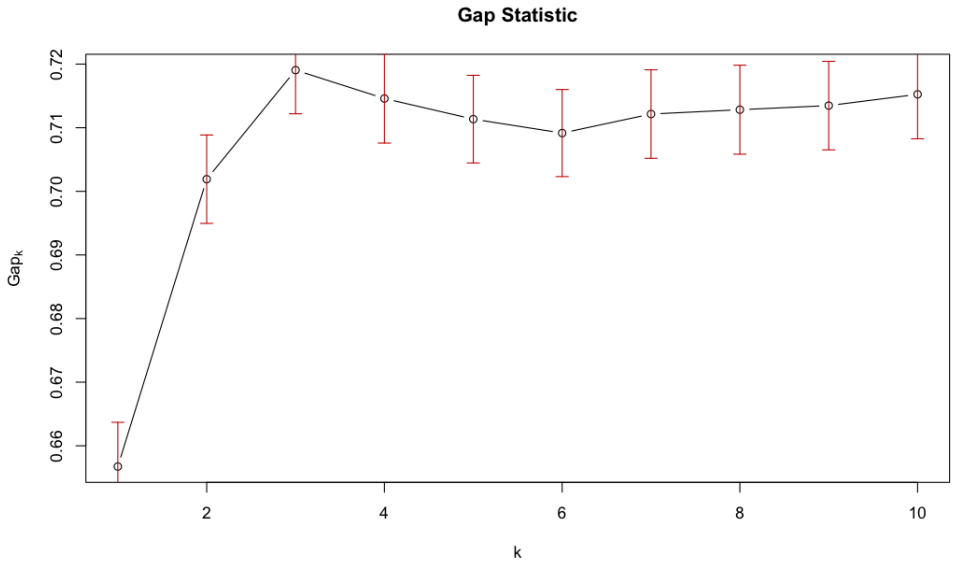

Fig S16: Gap Statistic for Dendritic Cells

References

[1] Amir et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology 31 545-552*, 2013.

[2] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS 97(18) 10101-10106*, 2000.

[3] Moon et al. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnology 37(12):1482-1492*, 2019.

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research 9:2579 – 2605*, 2008.

[5] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint arXiv:1802.03426v3*, 2020.

[6] Becht et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology 37 38-44*, 2019.

[7] Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology 39 156-157*, 2021.

[8] Francesco Crecchi, Cyril de Bodt, Michel Verleysen, John A. Lee, and Davide Bacciu. Perplexity-free parametric t-SNE. *arXiv preprint arXiv:2010.01359v1*, 2020.

[9] Haiyang Huang, Yingfan Wang, Cynthia Rudin, and Edward P. Browne. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Communications Biology, 5:716*, 2022.

[10] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications, 10:5416*, 2019.

[11] Yanshuai Cao and Luyu Wang. Automatic selection of t-SNE perplexity. *arXiv preprint arXiv:1708.03229.v1*, 2017.

[12] Ronald R. Coifman and Stéphane Lagon. Diffusion maps. *Applied and Computational Harmonic Analysis 21:1 5-30*, 2006.

[13] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to Use t-SNE Effectively. *Distill*, 2016.

[14] Andy Coenen and Adam Pearce for Google PAIR. Understanding UMAP. *https://pair-code.github.io/understanding-umap/*.

[15] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Computational Biology 19(8):e1011288*, 2023.

[16] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Towards a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics 27:3*, 2021.

[17] Jarkko Venna and Samuel Kaski. Visualizing gene interaction graphs with local multidimensional scaling. *European Symposium on Artificial Neural Networks*, 2006.

[18] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research 22*, 2021.

[19] Jesse H. Krijthe. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. *https://github.com/jkrijthe/Rtsne*, 2015.

[20] Po-Yuan Tung, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard, et al. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports 7:39921*, 2017.

[21] Dara M. Strauss-Albee, Julia Fukuyama, Emily C. Liang, Yi Yao, Justin A. Jarrell, Alison L. Drake, et al. Human NK cell repertoire diversity reflects immune experience and correlates with viral susceptibility. *Science Translational Medicine 7:297*, 2015.

[22] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, et al. Enterotypes of the human gut microbiome. *Nature 473 174-180*, 2011.

[23] Horn, John L. A rationale and test for the number of factors in factor analysis. *Psychometrika 30:2 179-185*, 1965.

[24] Martin Skrodzki, Nicolas Chaves-de-Plaza, Klaus Hildebrandt, Thomas Höllt, and Elmar Eisemann. Tuning the perplexity for and computing sampling-based t-SNE embeddings. *arXiv preprint arXiv:2308.15513v1*, 2023.

[25] Cell Ranger ARC 2.0.0. Single Cell Multiome ATAC + Gene Expression Dataset. *https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-3-k-1-standard-2-0-0*, 2021.

[26] Benjamin Parks. BPCells: Single Cell Counts Matrices to PCA. *https://bnprks.github.io/BPCells/articles/pbmc3k.html*, 2023.

[27] Shih-Kai Chu, Shilin Zhao, Yu Shyr, and Qi liu. Comprehensive evaluation of noise reduction methods for single-cell RNA sequencing data. *Briefings in Bioinformatics 23:2*, 2022.

[28] Ehsan Amid and Manfred K. Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204v2*, 2022.

[29] John A. Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing 72:1431 – 1443*, 2009.

[30] Tobias Schreck, Tatiana von Landesberger, and Sebastian Bremm. Techniques for precision-based visual analysis of projected data. *Sage 9:3*, 2012.