Supporting Information

SI Simulated Examples

SI.1 Trefoil Plots

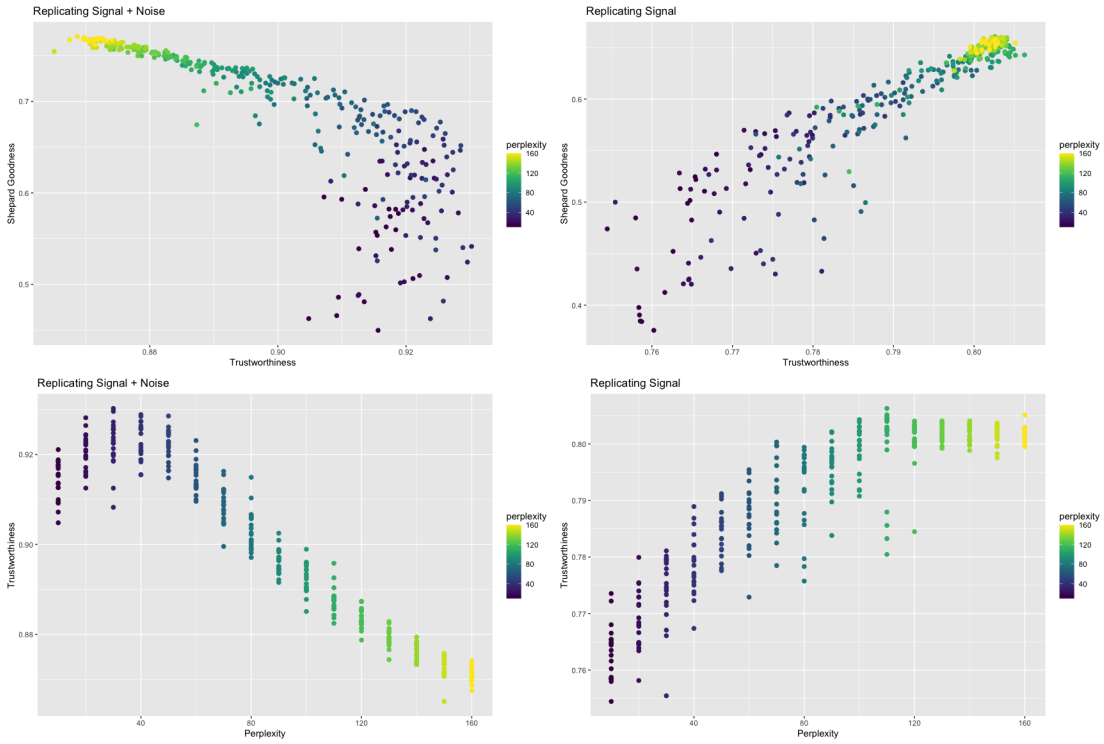

Fig S1: Trefoil Plots

SI.2 Mammoth Plots

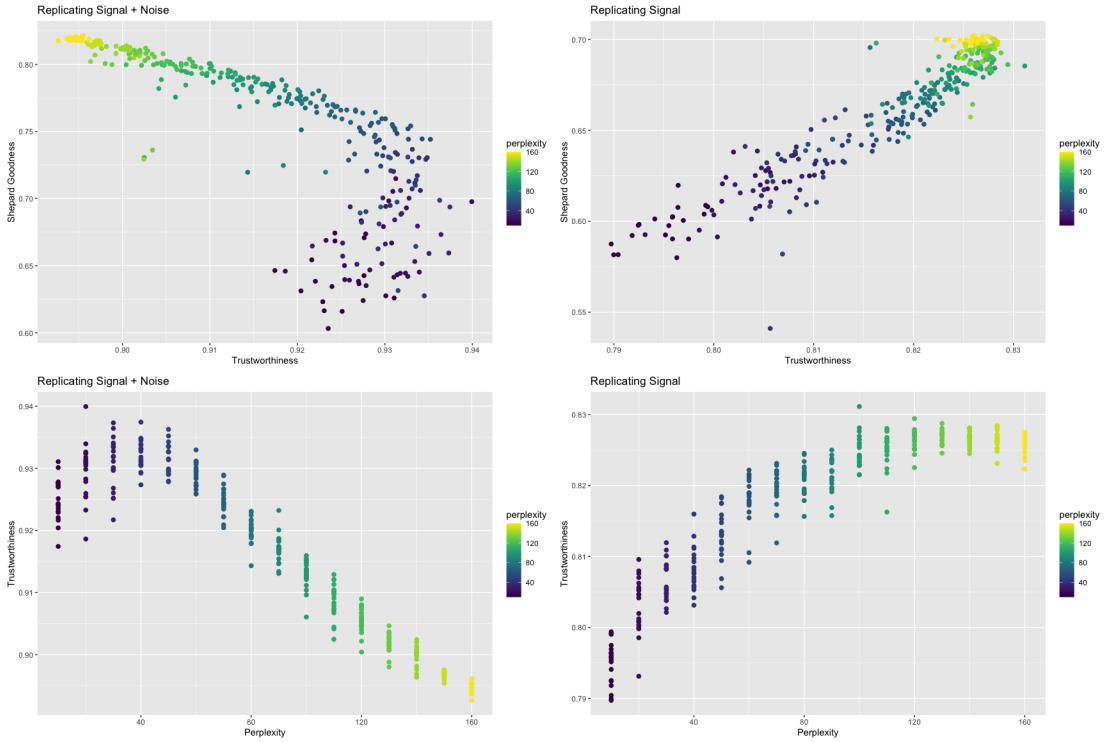

Fig S2: Mammoth Plots

SII Practical Examples

SII.1 UMAP Plots (CyTOF)

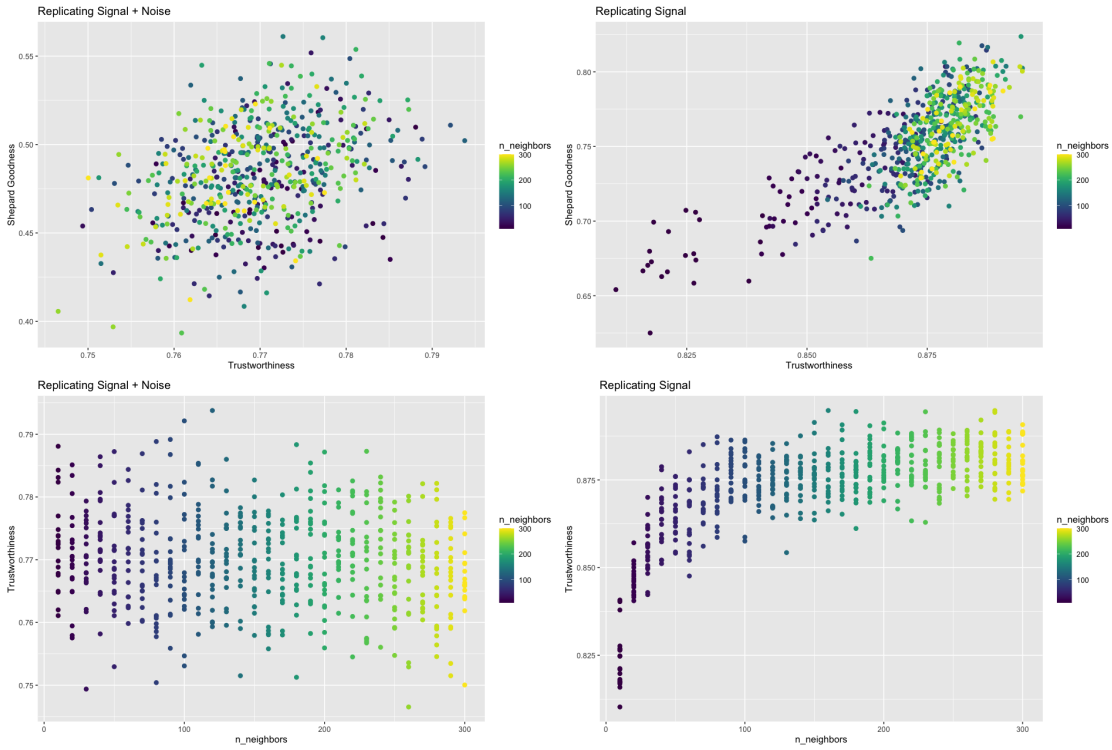

Fig S3: UMAP Plots

SII.2 scRNA Dataset

This is a dataset of induced pluripotent stem cells generated from three different individuals [13]. The original data includes 864 units and 19,027 readings per unit. To process this zero-inflated count data, columns containing a large proportion of 0's (20% or more) were removed before a log transformation was applied. This reduced the dimension to 5,431. A PCA pre-processing stop further reduced the dimension to 500, which still retained 88% of the variance of the log-transformed data. The signal was first taken to be the first five principal components, then the first 10 principal components. Notice the optimal perplexity when compared against the original data differed between these two experiments, even though it should theoretically be independent of the chosen signal dimension. This is due to the inherent randomness of the t-SNE algorithm.

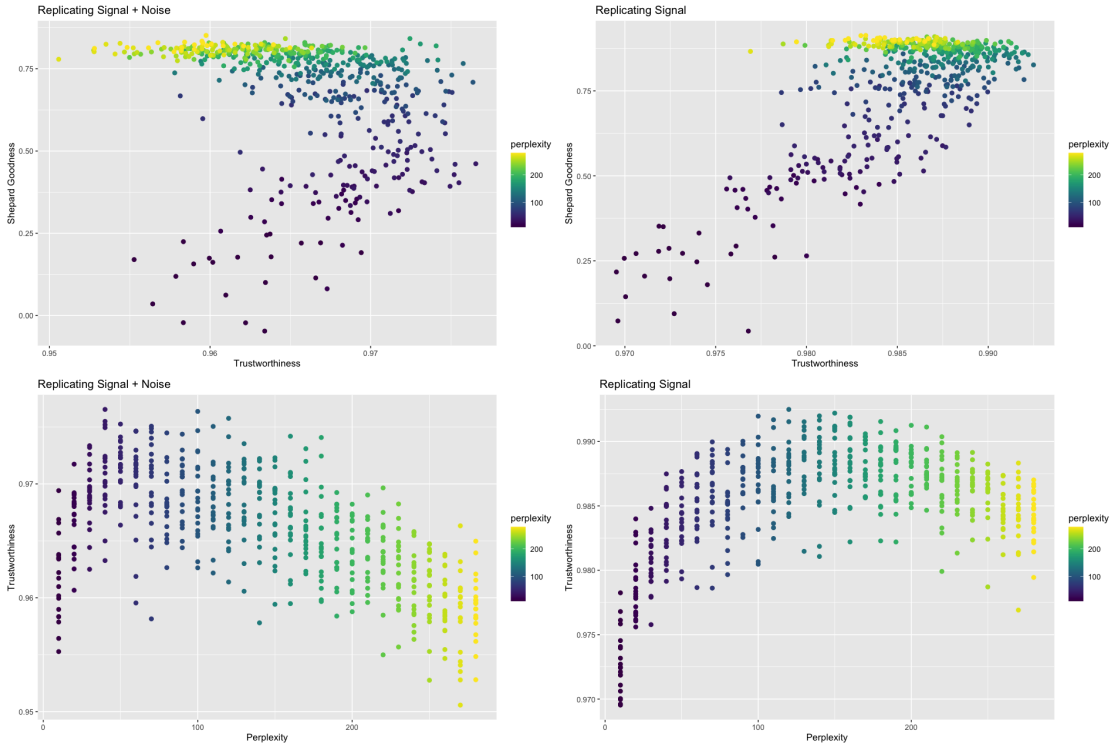

Fig S4: scRNA Plots (r = 5)

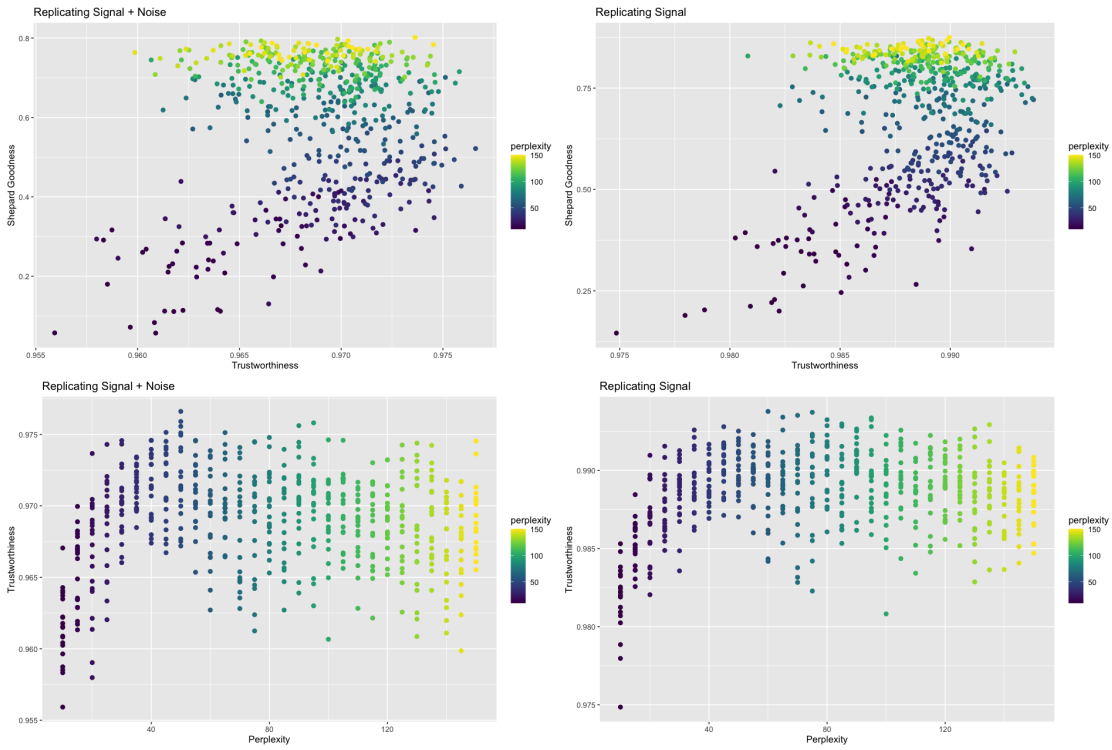

Fig S5: scRNA Plots (r = 10)

SII.3 Microbiome Dataset

[14] 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. Notice the optimal perplexity when compared against the original data differed between these two experiments, even though it should theoretically be independent of the chosen signal dimension. This is due to the inherent randomness of the t-SNE algorithm.

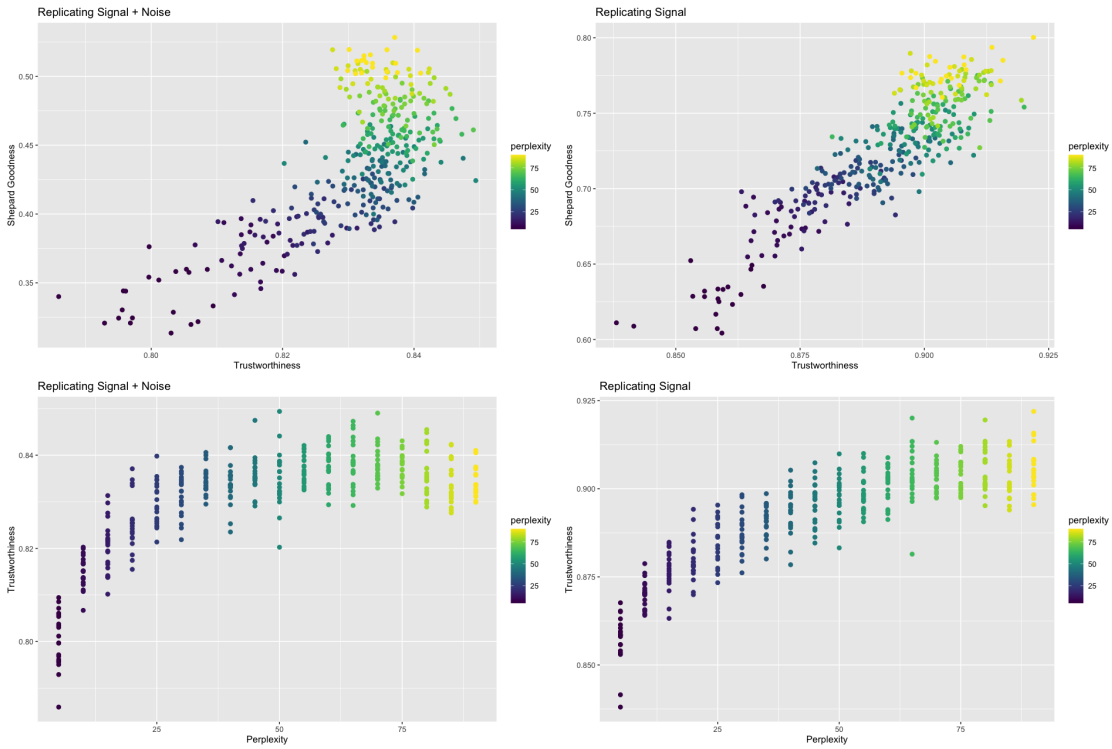

Fig S6: Microbiome Plots (r = 5)

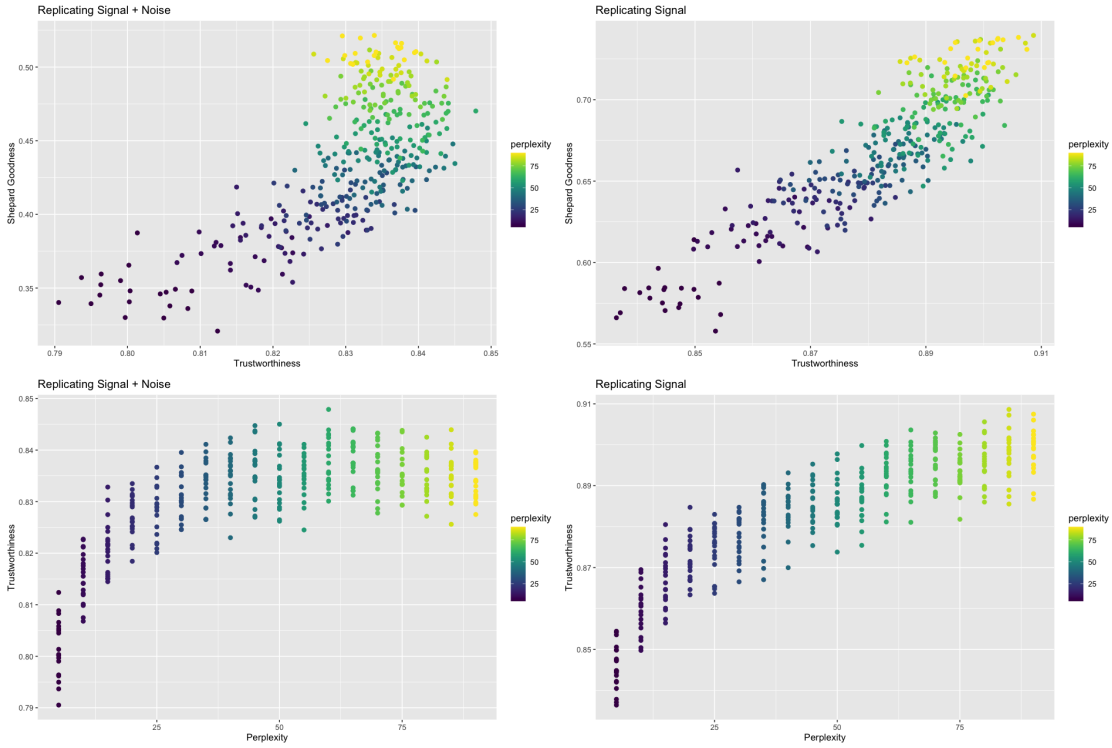

Fig S7: Microbiome Plots (r = 8)

References

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

[2] 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.

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

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

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

[6] Leland McInnes, John Healy, and James Melville. *arXiv preprint arXiv:1802.03426v3*, 2020.

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

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

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

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