MATTERS ARISING
https://doi.org/10.1038/s41587-020-00809-z

nature biotechnology

Initialization is critical for preserving global data structure in both *t*-SNE and UMAP

Dmitry Kobak [1] ✉ and George C. Linderman [2] ✉

ARISING FROM Becht, E. et al. *Nature Biotechnology* https://doi.org/10.1038/nbt.4314 (2019)

One of the most ubiquitous analysis tools in single-cell transcriptomics and cytometry is *t*-distributed stochastic neighbor embedding (*t*-SNE)[1], which is used to visualize individual cells as points on a two-dimensional scatterplot such that similar cells are positioned close together[2]. A related algorithm, called uniform manifold approximation and projection (UMAP)[3], has attracted substantial attention in the single-cell community[4]. In *Nature Biotechnology*, Becht et al.[4] argued that UMAP is preferable to *t*-SNE because it better preserves the global structure of the data and is more consistent across runs. Here we show that this alleged superiority of UMAP can be entirely attributed to different choices of initialization in the implementations used by Becht et al.: the *t*-SNE implementations by default used random initialization, while the UMAP implementation used a technique called Laplacian eigenmaps (LE)[5] to initialize the embedding. We show that UMAP with random initialization preserves global structure as poorly as *t*-SNE with random initialization, while *t*-SNE with informative initialization performs as well as UMAP with informative initialization. On the basis of these observations, we argue that there is currently no evidence that the UMAP algorithm per se has any advantage over *t*-SNE in terms of preserving global structure. We also contend that these algorithms should always use informative initialization by default.

At the core of both *t*-SNE and UMAP are loss functions that make similar points attract each other and push dissimilar points away from each other. Both algorithms minimize their loss functions by using gradient descent. Gradient descent begins with some initial configuration of points, and with each iteration the points are moved to decrease the loss function. The specific implementations of these algorithms used by Becht et al. differed in how the initial configuration of points was chosen: the *t*-SNE implementations placed the points randomly, whereas the UMAP implementation used LE[5], an algorithm that can often achieve globally accurate embedding on its own[6]. The effect of this difference on how well the two algorithms preserve global structure was not discussed or investigated by Becht et al.

We can illustrate the importance of initialization for both algorithms using a simple toy dataset (Fig. 1). We sampled $n=7{,}000$ points from a circle with some added Gaussian noise and used UMAP and *t*-SNE to construct an embedding. We kept all parameters for both algorithms at their default values and only changed the initialization. The *t*-SNE algorithm with random initialization produced a knot; UMAP with random initialization produced a tangled web with multiple tears. In both cases, the global structure was evidently not preserved; indeed, gradient descent in *t*-SNE and UMAP only pulls close neighbors together and is not much influenced by the global arrangement of points. At the same time, LE recovers the original circle for this toy dataset and, when used for initialization, strongly improves the UMAP result. A method often recommended for initialization in *t*-SNE is principal component analysis (PCA)[2]. Here PCA also recovers the original circle and strongly improves the *t*-SNE result. In both cases, only with informative initialization can UMAP and *t*-SNE produce a faithful representation of the closed one-dimensional manifold—the circle.

Using the code published by Becht et al., we analyzed the separate effects of initialization and algorithm on their results by adding UMAP with random initialization and *t*-SNE (using FIt-SNE[7]) with PCA initialization to the benchmarking comparison. Apart from the initialization, both algorithms were run with the same parameters as in Becht et al. We used all three datasets analyzed in the original publication (sample sizes from 320,000 to 820,000 cells)[8–10]. To quantify preservation of global structure, Becht et al. computed Pearson correlation between pairwise Euclidean distances in high-dimensional space and in the embedding. To quantify the reproducibility of the embedding, the authors embedded random subsamples of the data and measured the correlation of the coordinates of subsample embeddings with the coordinates of the full-dataset embeddings (up to symmetries around the coordinate axes). Our results show that *t*-SNE and UMAP with random initialization perform similarly poorly with regard to both metrics, whereas *t*-SNE and UMAP with PCA and LE initialization, respectively, perform similarly well (Table 1). See Extended Data Figs. 1–6 for the exact analogs of the original figures from Becht et al.

Becht et al. wrote that their findings were "consistent with the idea" that UMAP performs "optimizations that are sensitive to global features of the data, thus reaching similar arrangements more consistently." Our results show that this conclusion can be misleading: their findings were in fact not due to UMAP optimizations but rather to its initialization.

We have recently argued that, for single-cell transcriptomic data, *t*-SNE with PCA initialization produces more meaningful embeddings than *t*-SNE with random initialization[2]. The findings of Becht et al., when interpreted correctly, also underscore the importance of using informative initialization and suggest that it should be used as the default option in *t*-SNE and UMAP implementations. PCA initialization has always been the default in openTSNE[11], a Python reimplementation of FIt-SNE, and FIt-SNE v.1.2 now also uses it by default. OpenTSNE v.0.4 now also supports LE initialization. Importantly, *t*-SNE with non-random initialization should not be considered a new algorithm or even an extension of the original *t*-SNE; it is exactly the same algorithm with the same loss function, and almost any existing implementation trivially allows the use of any given initialization, including the PCA-based one.

[1]Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany. [2]Applied Mathematics Program, Yale University, New Haven, CT, USA.
✉e-mail: dmitry.kobak@uni-tuebingen.de; george.linderman@yale.edu

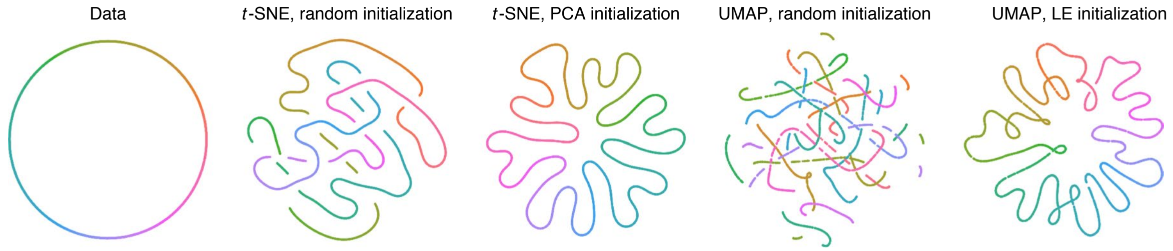


Fig. 1 | *t*-SNE and UMAP with random and non-random initialization. Embeddings of $n = 7{,}000$ points sampled from a circle with a small amount of Gaussian noise ($\sigma = r/1{,}000$, where r is the circle's radius). We used random and PCA initialization for *t*-SNE (openTSNE[11] v.0.4.4) and random and LE initialization for UMAP (v.0.4.6). All other parameters were kept as default. For this dataset, PCA and LE give the same initialization. Note that openTSNE scales PCA initialization to have s.d. = 0.0001, which is the default s.d. for random initialization in *t*-SNE[2]; similarly, UMAP scales the LE result to have a span of 20, which is the value it uses for random initialization.

Table 1 | Performance of *t*-SNE and UMAP with random and informative initialization using datasets and evaluation metrics from Becht et al.

	Preservation of pairwise distances			Reproducibility of large-scale structures[a]		
Dataset	**Samusik et al.[8]**	**Wong et al.[9]**	**Han et al.[10]**	**Samusik et al.[8]**	**Wong et al.[9]**	**Han et al.[10]**
UMAP, LE initialization	**0.70**	0.57	**0.30**	0.94	**0.98**	0.49
UMAP, random initialization	0.41	0.38	0.14	0.24	0.21	0.22
t-SNE, PCA initialization	0.59	**0.66**	0.28	**0.95**	**0.98**	**0.92**
t-SNE, random initialization	0.32	0.36	0.18	0.29	0.33	0.06

[a] For the reproducibility metric, the average over three random subsamples of size $n = 200{,}000$ is reported. Bold numbers denote the maximum value in each column.

Our aim here was not to argue which algorithm, *t*-SNE or UMAP, is more suitable for single-cell studies. Once informative initialization is used, the two algorithms appear to preserve the global structure similarly well, and modern implementations of the two algorithms work with similar speed (the widespread opinion that UMAP is much faster than *t*-SNE is outdated: for two-dimensional embeddings, FIt-SNE works at least as quickly[2,7]). When comparing the resulting embeddings (Extended Data Figs. 3–6), the most striking difference is that UMAP produces denser, more compact clusters than *t*-SNE, with more white space in between. Very similar embeddings can be produced by *t*-SNE with so-called exaggeration, which increases the attractive forces by a constant factor[2]. Future research in machine learning is needed to pinpoint the exact mathematical and algorithmic origins of this difference[12], and future research in single-cell biology is needed to decide which algorithm is more faithful to the underlying biological data. It remains challenging to measure how faithful a given embedding is, and we believe this is an important topic for future work in data visualization for single-cell biology and beyond.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-00809-z.

Received: 2 December 2019; Accepted: 23 December 2020;
Published online: 1 February 2021

References

1. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
2. Kobak, D. & Berens, P. The art of using *t*-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
3. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
4. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–40 (2019).
5. Belkin, M. & Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems* 585–591 (2002).
6. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
7. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast interpolation-based *t*-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **16**, 243–245 (2019).
8. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).
9. Wong, M. T. et al. A high-dimensional atlas of human T cell diversity reveals tissue-specific trafficking and cytokine signatures. *Immunity* **45**, 442–456 (2016).
10. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).
11. Policar, P. G., Strazar, M. & Zupan, B. openTSNE: a modular Python library for *t*-SNE dimensionality reduction and embedding. Preprint at *bioRxiv* https://doi.org/10.1101/731877 (2019).
12. Böhm, J. N., Berens, B. & Kobak, D. A unifying perspective on neighbor embeddings along the attraction–repulsion spectrum. Preprint at https://arxiv.org/abs/2007.08902 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Data availability

The data in this study were sourced from refs. [8–10].

Code availability

The R code extending the analysis of Becht et al. is available at https://github.com/linqiaozhi/DR_benchmark_initialization. The Python code used to produce Fig. 1 is available at https://github.com/dkobak/tsne-umap-init.

Acknowledgements

The authors thank P. Berens, S. Steinerberger and Y. Kluger for discussions and helpful comments. D.K. was supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1 and the Cluster of Excellence 'Machine Learning—New Perspectives for Science', EXC 2064, project number 390727645), the Federal Ministry of Education and Research (FKZ 01GQ1601 and 01IS18039A) and the National Institute of Mental Health of the National Institutes of Health under award number U19MH114830. G.C.L. was supported by the National Human Genome Research Institute (F30HG010102) and National Institutes of Health MSTP training grant T32GM007205. A portion of the benchmarks were run on computational resources funded by the National Institutes of Health (R01GM131642, principal investigator: Y. Kluger). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

The authors contributed equally.

Competing interests

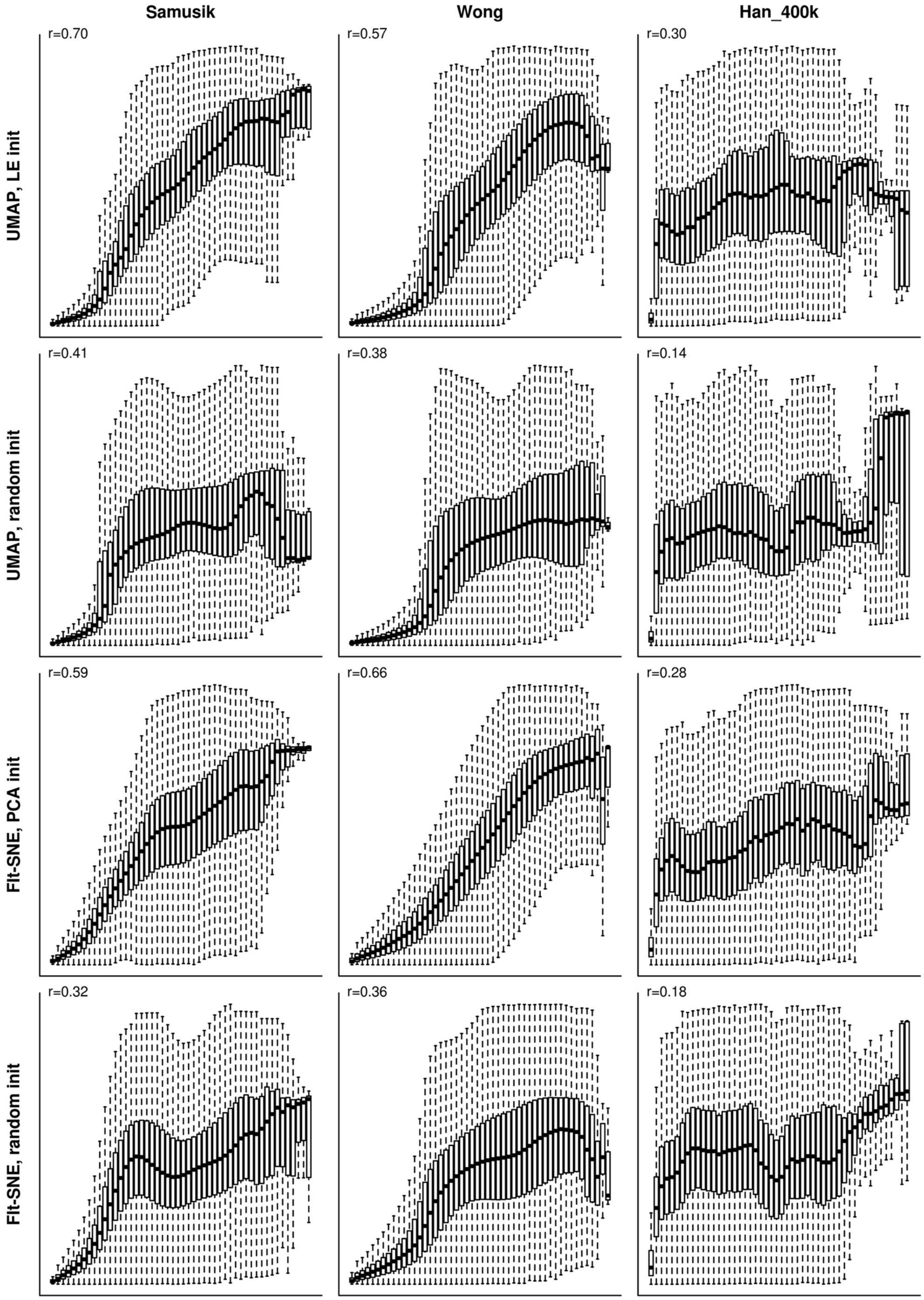
The authors declare no competing interests.

Additional information

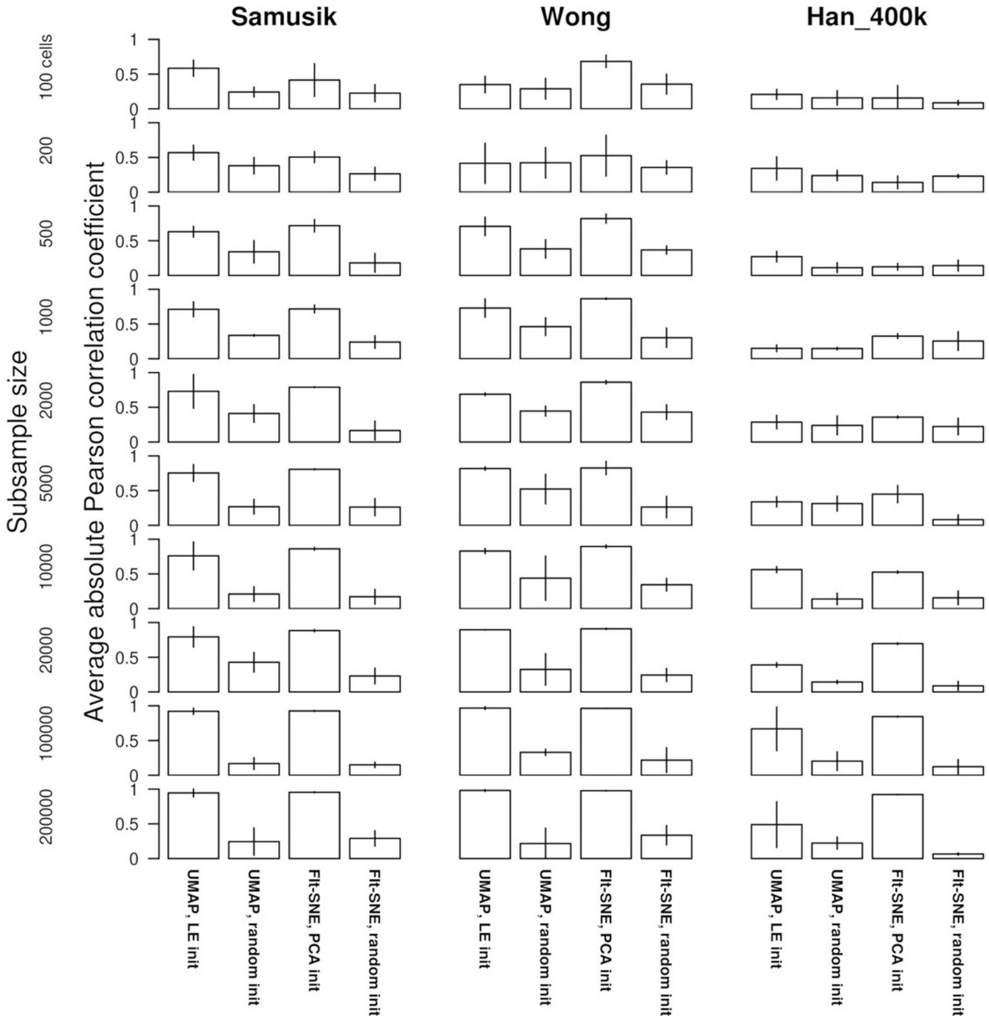
Extended data is available for this paper at https://doi.org/10.1038/s41587-020-00809-z.

Correspondence and requests for materials should be addressed to D.K. or G.C.L.

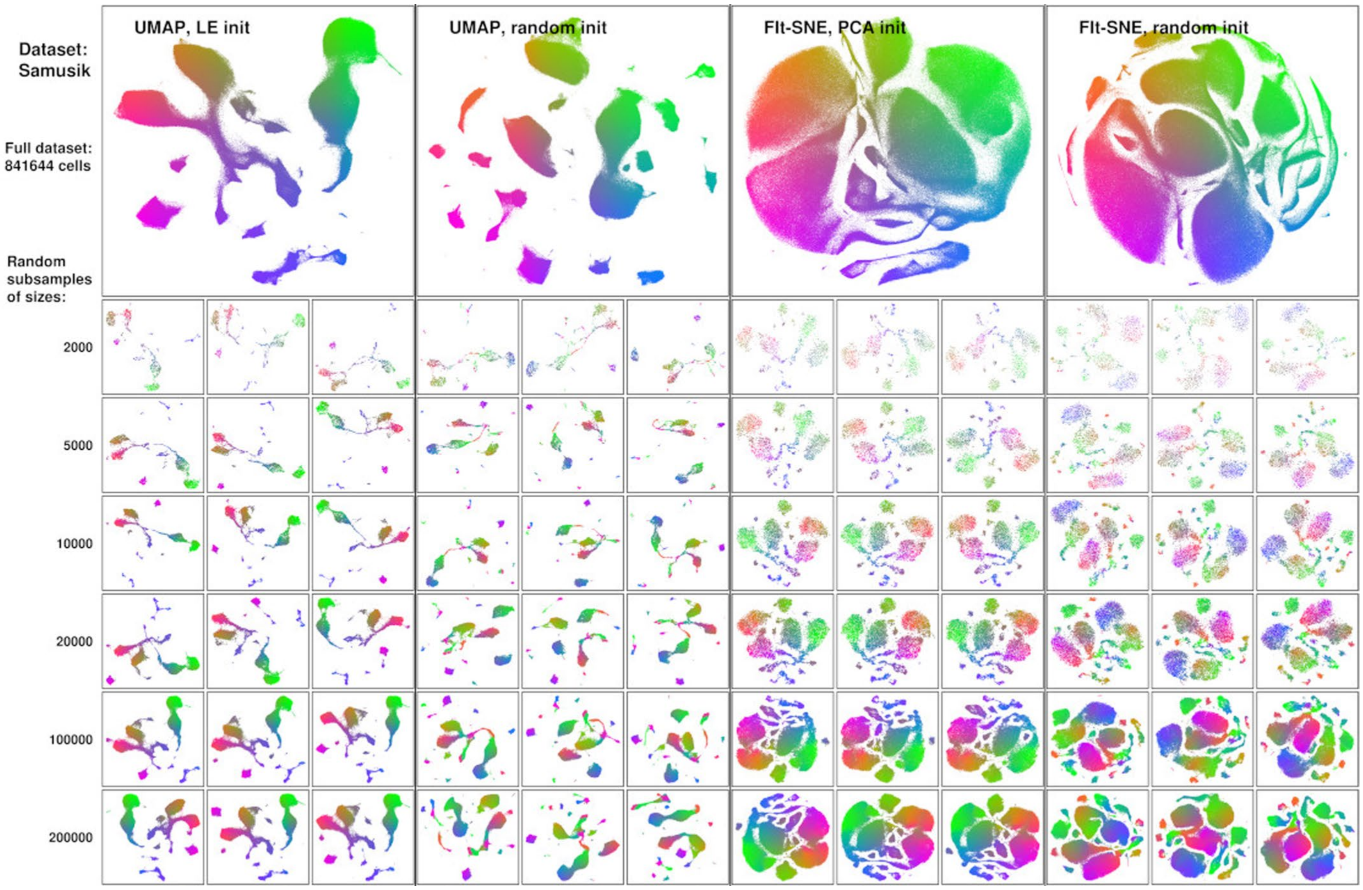
Reprints and permissions information is available at www.nature.com/reprints.



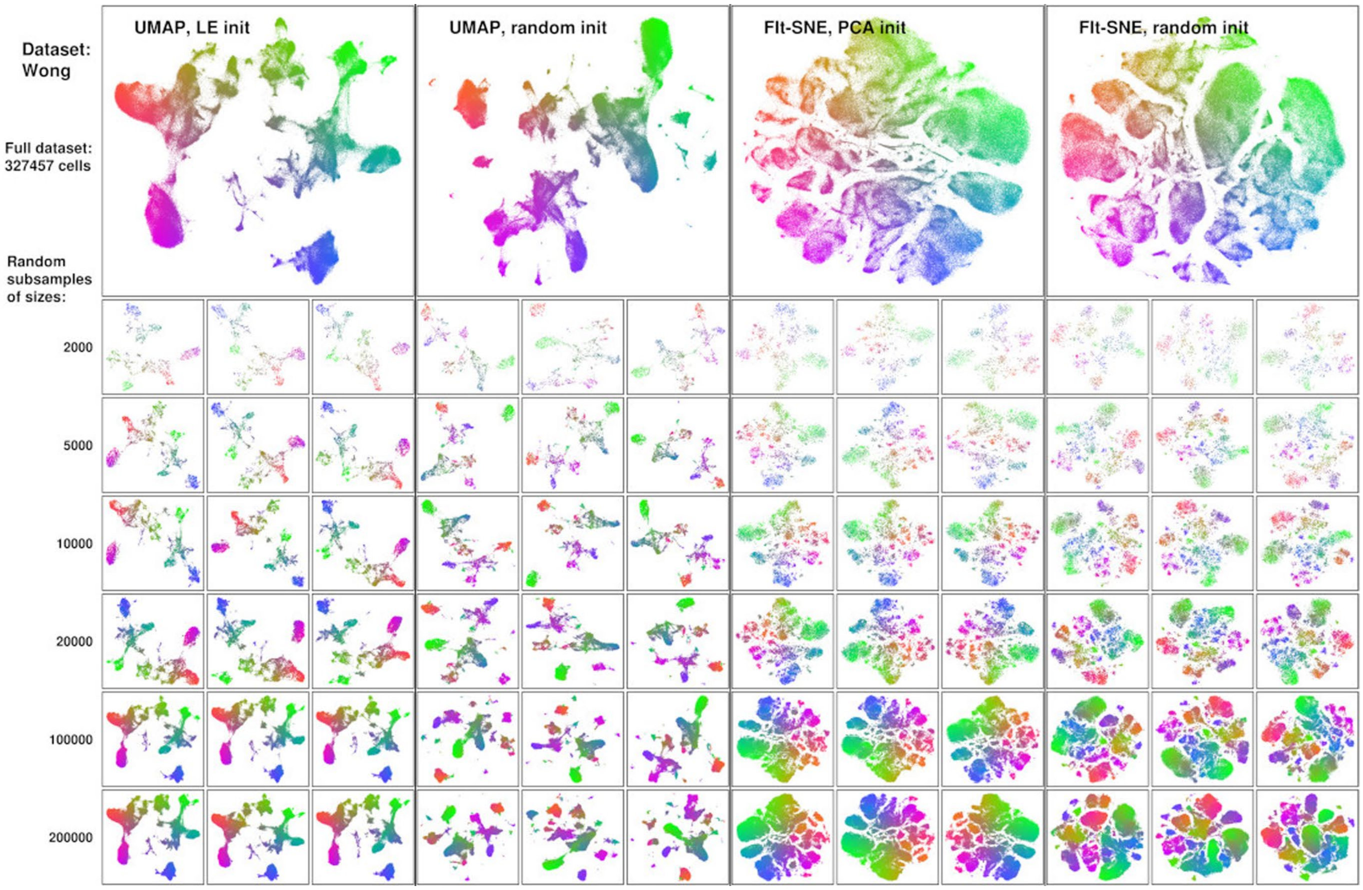
Extended Data Fig. 1 | Preservation of pairwise distances in embeddings. The exact analogue of Fig. 5 in the original publication by Becht et al.[4] To quote the original caption: 'Box plots represent distances across pairs of points in the embeddings, binned using 50 equal-width bins over the pairwise distances in the original space using 10,000 randomly selected points, leading to 49,995,000 pairs of pairwise distances. […] The value of the Pearson correlation coefficient computed over the pairs of pairwise distances is reported. For the box plots, the central bar represents the median, and the top and bottom boundary of the boxes represent the 75th and 25th percentiles, respectively. The whiskers represent 1.5 times the interquartile range above (or, respectively, below) the top (or, respectively, bottom) box boundary, truncated to the data range if applicable.' We recomputed all embeddings (except for the UMAP with LE initialization of the Wong et al.[9] dataset, which was loaded from external source, as in the code accompanying the original publication). All algorithms were run with the same parameters as in the original publication (which always were the default parameters, apart from n_neighbors set to 30 in UMAP for the Han et al.[10] dataset; we kept this value for both initializations). We used the same version of FIt-SNE as in the original publication, to make sure that all the default parameters stayed the same. Y-axis goes from zero to the maximum pairwise distance in all subplots.



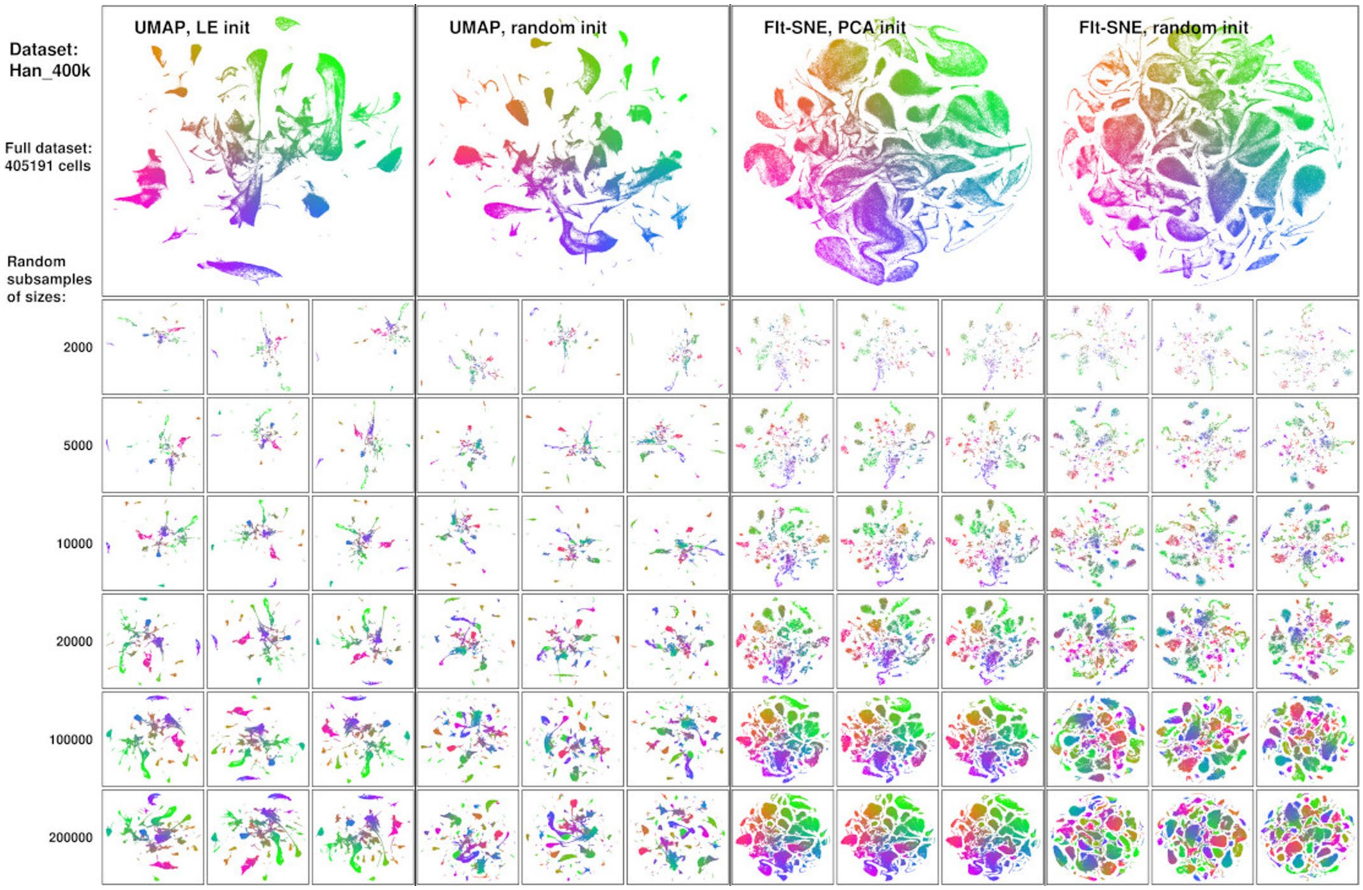
Extended Data Fig. 2 | Reproducibility of large-scale structures in embeddings. The exact analogue of Fig. 6 in the original publication[4]. To quote the original caption: 'Bar plots represent the average unsigned Pearson correlation coefficient of the points' coordinates in the embedding of subsamples versus in the embedding of the full dataset, thus measuring the correlation of coordinates in subsamples versus in the embedding of the full dataset, up to symmetries along the graph axes. Bar heights represent the average across three replicates and vertical bars the corresponding s.d.'



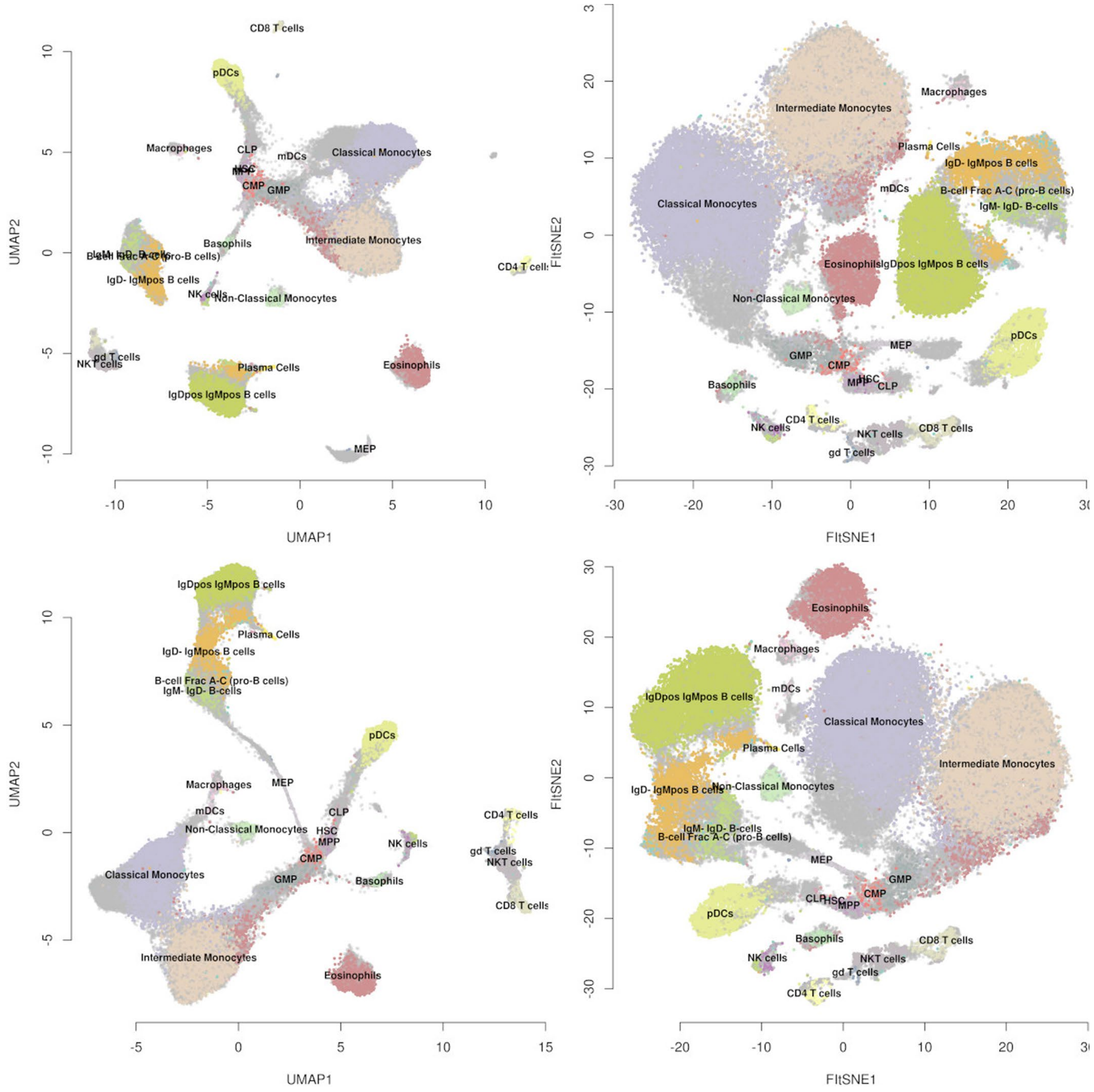
Extended Data Fig. 3 | Qualitative assessment of the reproducibility of embeddings using the Samusik et al.[8] dataset. The exact analogue of Supplementary Fig. 7a from the original publication[4]. To quote the original caption: 'Embeddings of full datasets as well as subsamples of varying sizes replicated thrice for [four] dimensionality reduction methods. The color-code is generated using the embedding of the full dataset and propagated to the subsamples.'



Extended Data Fig. 4 | Qualitative assessment of the reproducibility of embeddings using the Wong et al.[9] dataset. The exact analogue of Supplementary Fig. 7b from the original publication[4]. To quote the original caption: 'Embeddings of full datasets as well as subsamples of varying sizes replicated thrice for [four] dimensionality reduction methods. The color-code is generated using the embedding of the full dataset and propagated to the subsamples.'



Extended Data Fig. 5 | Qualitative assessment of the reproducibility of embeddings using the Han et al.[10] dataset. The exact analogue of Supplementary Fig. 7c from the original publication[4]. To quote the original caption: 'Embeddings of full datasets as well as subsamples of varying sizes replicated thrice for [four] dimensionality reduction methods. The color-code is generated using the embedding of the full dataset and propagated to the subsamples.'



Extended Data Fig. 6 | Annotated embeddings of the Samusik_01 dataset (sample size n=86,864). Top row: UMAP with random initialization (left) and t-SNE with random initialization (right). Bottom row: UMAP with default initialization (left) and t-SNE with PCA initialization (right). The bottom-left and upper-right panels are analogues of Fig. 2a,b from the original publication[4]. Note that the T cells are not colocalized in the UMAP embedding with random initialization.