# Exploring dimension-reduced embeddings with Sleepwalk

Svetlana Ovchinnikova and Simon Anders*
*Center for Molecular Biology (ZMBH), University of Heidelberg, Germany*
(Dated: 2019-04-01)

Abstract

Dimension-reduction methods, such as t-SNE or UMAP, are widely used when exploring high-dimensional data describing many entities, e.g., RNA-Seq data for many single cells. However, dimension reduction is unavoidably prone to introducing artefacts, and we hence need means to see where a dimension-reduced embedding is a faithful representation of the local neighbourhood and where it is not.

We present Sleepwalk, a simple but powerful tool that allows the user to interactively explore an embedding, using colour to depict "true" similarities of all points to the cell under the mouse cursor. We show how this approach not only highlights distortions, but also reveals otherwise hidden characteristics of the data, and how Sleepwalk's comparative modes help integrate multi-sample data and understand differences between embedding and preprocessing methods. Sleepwalk is a versatile and intuitive tool that unlocks the full power of dimension reduction and will be of value not only in single-cell RNA-Seq but also in any other area with matrix-shaped big data.

## I. INTRODUCTION

Whenever one is presented with large amounts of data, producing a suitable plot to get an overview is an important first step. So-called dimension reduction methods are commonly used. For example, in high-throughput transcriptomics projects using expression microarrays or RNA-Seq, it is common practice, especially when working with many samples, to perform principal component analysis (PCA) on a suitably normalized and transformed expression matrix and then plot the samples' first two principal components as a scatter plot. Of course, PCA has more uses than just providing such an overview plot (See Ref [1] for a primer.), but nevertheless, the user's expectation is often simply that samples with similar expression profile should appear close together ("cluster together"), while samples with strong differences should appear farther apart. PCA's popularity in biology notwithstanding, the literature offers many methods designed specifically with this goal in mind, with the best-known classic example perhaps being classical multidimensional scaling (classical MDS, also known as principal coordinate analysis, PCoA), Kruskall's non-metric multidimensional scaling [2] and Kohonen's self-organizing maps (SOM) [3].

The recent rapid progress of single-cell RNA-Seq methods, now enabling the measurement of expression profiles of thousands of individual cells in a sample, has renewed biologists' interest in dimension reduction methods. Here, t-distributed stochastic neighbour embedding (t-SNE, [7], Figure 1) has become a de-facto standard, with Uniform Manifold Approximation and Projection (UMAP, [8]) currently gaining popularity as an alternative. Other dimension reduction methods popular in the single-cell RNA-Seq field include diffusion maps [9, 10], the Monocle methods [11, 12], DDRTree [13] and more.

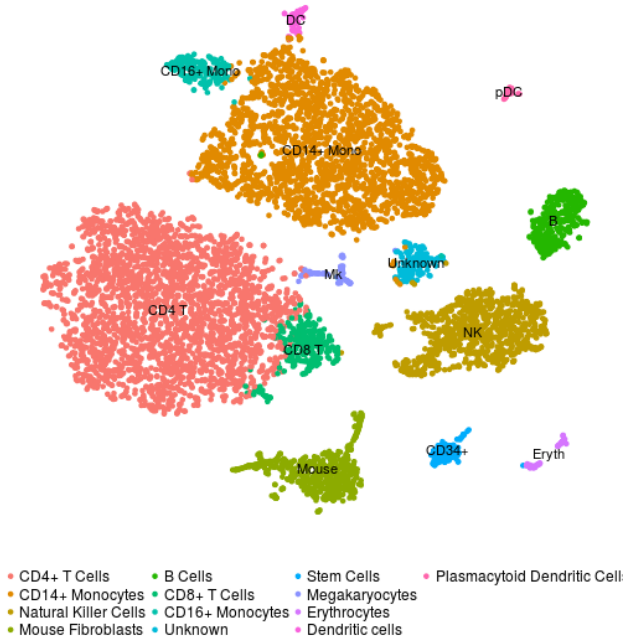



FIG. 1: Example of a t-SNE plot: These are cord-blood mononuclear cells studied by Ref. [4]. The embedding and the assignment of cell types have been taken from the Seurat [5] tutorial that uses this data set as example [6].

These varied methods have been developed with different design goals: for example, some methods strive to primarily preserve neighborhood, others to represent the overall structure or larger-scale relations. Nevertheless, when using any of them to gain an overview over some data, the practitioner's primary expectation is usually that cells depicted close to each other or within the same apparent structure or cluster have more similar expression profiles than cells depicted in different regions of the plot or in different structures. It is, however, impossible to provide

*Electronic address: sanders@fs.tum.de

a rendition that represents all the distances between expression profiles faithfully, without any distortions, unless the data was on a two-dimensional plane in the first place. (The impossibility to provide a distortion-free map of the Earth on a flat sheet of paper is a good example for this.)

As distortions are unavoidable, the viewer of a dimension-reduced embedding should be made aware of them. We need a means to see where in a given 2D embedding the representation is faithful to our expectations and where it is not.

## II. RESULTS

### A. The Sleepwalk app

Here, we present "sleepwalk", an interactive tool that provides an intuitive solution to the task just outlined.

It works as follows: The user provides an embedding, i.e., the two-dimensional coordinates output by a dimension-reducing method, and a matrix of the original distances between all the points that the embedding has tried to (but cannot entirely have succeeded to) preserve. Sleepwalk displays the embedding on the screen, and whenever the user hovers with the mouse over a data point, all points are coloured according to their distance to the focus point under the mouse cursor (Figure 2). By moving the mouse over the plot, the user can explore how the faithfulness of the embedding varies between regions of the plot. Buttons are provided to change the range of the colour scale.

We invite the reader to pause here and try Sleepwalk himself or herself. At https://anders-biostat.github.io/sleepwalk/supplementary/ (and also in this paper's supplement HTML file), life interactive sleepwalk renditions of Figure 2 (and also of all the subsequent figures discussed below) are given. The Sleepwalk app runs in any Javascript-enabled web browser, i.e., it suffices to open the page or file in a browser, without need to install anything.

### B. Exploring an embedding

Sleepwalk makes aspects visible that are not apparent from a dimension reduction alone. For example, the two large clusters under the cursor in Figure 2a and Figure 2b have quite different characteristic. In the T cell cluster (Figure 2a), most of the cells have equal distance to each other: the cluster shows up as a large green cloud no matter where one points the mouse. The monocyte cluster (Figure 2b), however, spreads over larger distances: only a part shows up in green, which "follows" the mouse. In a static t-SNE plot (such as Figure 1), this cannot be seen.

We can also check the cluster borders and discover, for instance (Figure 2c), that some cells in the monocyte cluster are more similar to those in the T cell cluster than to those in their own cluster. They may have been assigned the wrong cell type, or might be doublets. Thus, the Sleepwalk exploration can alert the analyst to the need for further investigation of possibly misleading features of a dimension-reduced embedding.

When switching the colour scale to a wider distance range, we can also see here how Sleepwalk allows to judge relationships *between* clusters (Figure 2d): we see which clusters are more and which are less similar to each other – an information that a static t-SNE does not show, due to the method's design focus on faithful representation only of neighbourhoods.

### C. Feature-space distances

The colours in Sleepwalk are meant to indicate similarity or dissimilarity between the cells' expression profiles, quantified as distances. There are multiple suggestions for useful distance measures in the literature, and the users can provide whichever they prefer. To produce the t-SNE embedding in Figure 1, we followed the Seurat workflow [6], which calculates distances in a specific manner (Methods), and these are then also used by the t-SNE routine. We have also used these same distances to colour the points in the Sleepwalk rendition (Figure 2); thus allowing us to see directly where t-SNE succeeded and where it failed in its design goal of preserving the neighborhood relation in its input data.

t-SNE uses a flexible approach to define the distance scale over which cells are considered neighbours: it adjusts the distance scale for each cell such that all cells have approximately the same number of neighbours (the so-called perplexity). Sleepwalk, in contrast, uses a fixed distance scale. This is on purpose: it allows us to note where the neighbourhood has longer or shorter range (as shown in the comparison of Figures 2a and 2b). The app offers two buttons to increase or decrease the scale of the distance-to-colour mapping, allowing the user to assess the embedding at several scales (Figure 2d).

This also highlights how clusters that are apparent in an embedding can cease to appear separated if one increases the distance scale. An important use case for Sleepwalk is hence to understand an obtained clustering, and to alert the user to cases where the same amount of dissimilarity may cause cells to be put into different clusters in one Figures 2a and 2b).

We hasten here to acknowledge that clustering should not be performed on 2D embeddings, but on the original high-dimensional data or a medium-dimensional intermediate. In our example data, Seurat has performed a shared nearest neighbour (SNN) modularity optimization based clustering [14] in the space spanned by the first 13 principal components (Figure 1). Using a 2D embedding as a map of such clusters will nevertheless help understand them, and Sleepwalk will help the user explore them (Figure 2).

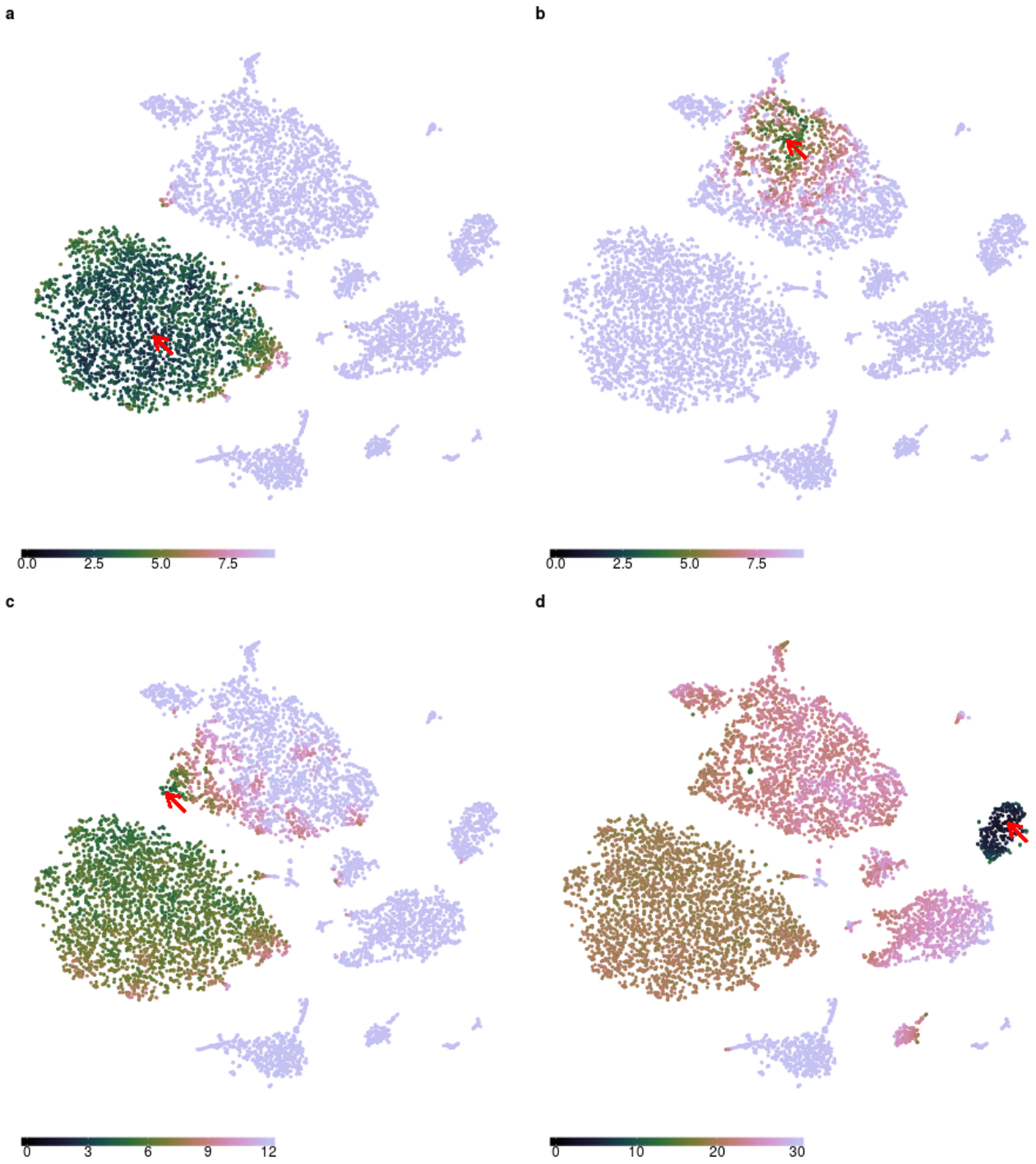



FIG. 2: The "Sleepwalk" app, being used to explore the t-SNE rendition of the cord-blood data set from Figure 1. The plots here are snapshots of a running "Sleepwalk" app; a life version can be found at https://anders-biostat.github.io/sleepwalk/supplementary/. The red arrow shows the current mouse position. By moving the mouse cursor through the embedding, we find, e.g., that the CD4+ T cell cluster is very tight and homogeneous (a), while the monocyte cluster shows much more heterogeneity (b), when comparing the colouring at the same colour scale. (c) Placing the mouse on this small tip of the monocyte cluster reveals that the cells there are more similar to the T cells than to the other monocytes, indicating that the cluster boundary might be inaccurate in both the t-SNE rendition and the SNN clustering on which the Seurat workflow's cell-type assignment is based. (d) With the colour scale set to a wider distance range, we can assess similarities *between* clusters: As expected, B cells are somewhat similar to T cells, less so to NK cells and monocytes, and distant to erythrocytes and the spiked-in mouse cells.

### D. Comparing embeddings

With the availability of choice in dimension reduction methods, the question arises which one to use. Benchmark comparisons may address this question in general; see for example [15] for a comparison of UMAP with t-SNE and related methods. When working on a specific dataset, however, simply calculating multiple embeddings and comparing them side by side might be even more helpful. We demonstrate this here using data from a study of the development of the mouse cerebellum [16]. In Figure 3, we show cells from development time point E13.5, first visualized with t-SNE (Figure 3a), then with UMAP (Figure 3b). The live apps in the HTML supplement show, in addition to this, also the same comparison for the cord-blood data of Figures 1 and 2.

To compare the two embedding, we need, at minimum, a way to see which points in the two plots correspond to the same cells. A classical approach is "brushing" [17]: selecting with the mouse a group of adjacently depicted cells in one plot causes them to be highlighted in the other one, too. Sleepwalk adapts this idea, but instead of the usual brush, we simply use in all embeddings the same colour for points corresponding to the same cell. Moving the mouse over points in one plot then highlights the neighbourhood structure induced by the feature-space distance chosen for that embedding not only there but also in all displayed embeddings and so links them. This allows us to see for a structure in one embedding whether there are corresponding structures in the other embeddings.

In the example shown in Figure 3, we see a clear correspondence between the major structures generated by t-SNE and by UMAP. Even the arrangement of cells within these structures is the same, which one can follow in the life version of the app. There are, however, also differences: The cells at the mouse position in Figure 3 are part of the connecting "filament" in the t-SNE embedding, but lie in an external "protrusion" in the UMAP.

### E. Comparing samples

Until recently, most single-cell RNA-Seq studies analysed only a single sample comprising many cells. Yet, the full value of the technique might become apparent only when it is used to compare between many samples. One currently popular approach to do so visually is to simply combine the data from the cells of all samples into one large expression matrix and perform t-SNE or UMAP on this. Often, global differences between samples, typically due to technical effects [18], will prevent similar cells from different samples to appear in the same cluster or structure in the dimension-reduced embedding. Methods to automatically remove such sample-to-sample differences (e.g., the CCA-based method in [5] and the MNN method in [19]) address this issue, but will not work always and may risk also removing biological signal.

An visual alternative is to produce a dimension-reduced embedding separately for each sample, and then try to find correspondences between the features in these. In Figure 4, we show how Sleepwalk allows to perform such an exploration comparing UMAP renderings for the two E13.5 and one of the E14.5 samples of the mouse cerebellum data set. Exploring the data with the mouse shows the two replicas of E13.5 samples (Figures 4a and 4b) are almost identical. The two branches (GABAergic and glutamatergic neurons) can be easily followed from the early progenitor cells up to the most differentiated ones. Comparing between the two E13.5 replicates reveals which aspects of the peculiar two-pronged shape of the glutamatergic branch is simply due to random variation and what seems reproducible. In the later E14.5 sample, the branches have disconnected from the progenitor cells, but Sleepwalk allows us to still identify corresponding cells. Interestingly, Sleepwalk can show that the GABAergic lineage is differentiated further in the E14.5 than in E13.5 samples, as the endpoint of the branch in E14.5 corresponds to an intermediate point in E13.5. Sleepwalk allows one to discover such details immediately, with minimal effort. Of course, such a visual exploration cannot replace a tailored detailed analysis but it is does provide a starting point and a first overview.

Crucially, using Sleepwalk's multi-sample comparison mode does not require any removal of global sample-to-sample differences with batch-effect correction methods. If the user selects a cell with the mouse in one sample, the cells that are similar to it will be highlighted, both in the same sample as well as in all other samples. This works even if the cells in the other samples seem much more distant, due to the additional sample-to-sample distance; we only might need to increase the scale of the distance-to-colour mapping for the cross-sample comparisons.

### F. Comparing distance metrics

In the examples discussed so far, we have used the feature-space distances as a kind of "ground truth" and used Sleepwalk to check whether a given embedding shows it well. However, there is no clearly best way to calculate the distance between two cells' expression profiles. For instance, we may either choose to use all genes in the calculation, or only some genes, which may either be chosen for having high expression and hence good signal, or perhaps chosen, via manual curation, as especially informative with respect to cell type or state. We may use the genes as they are or aggregate them before into meta- or eigen-genes, e.g., by a principal component analysis. We may use Euclidean distance, correlation distance or other measures. And, last but not least, the way how the expression data has been transformed, normalized or preprocessed can be understood as part of the choice of distance metric. Therefore, it is helpful to be able to compare different distance metrics.

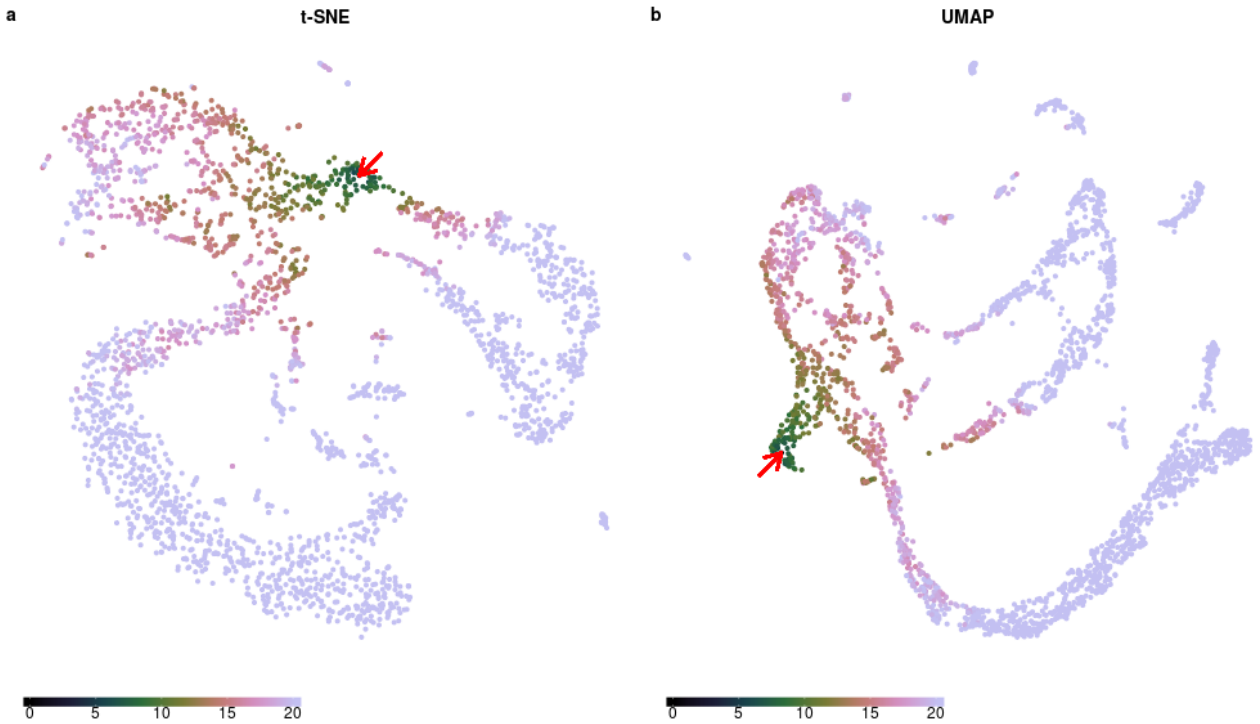



FIG. 3: Sleepwalk being used to compare two embeddings of the same single-cell data of a developing murine cerebellum at embryonal time point E13.5 [16]: t-SNE on the left and UMAP on the right. The user can explore one embedding in the same ways as in Figure 2, while all other embeddings that are displayed concurrently, are "slaved" to the one under the mouse cursor: each cell has the same colour in all embeddings. The red arrow shows the current mouse position. A life version (for this data and for the cord-blood data) can be found on the supplementary web page.

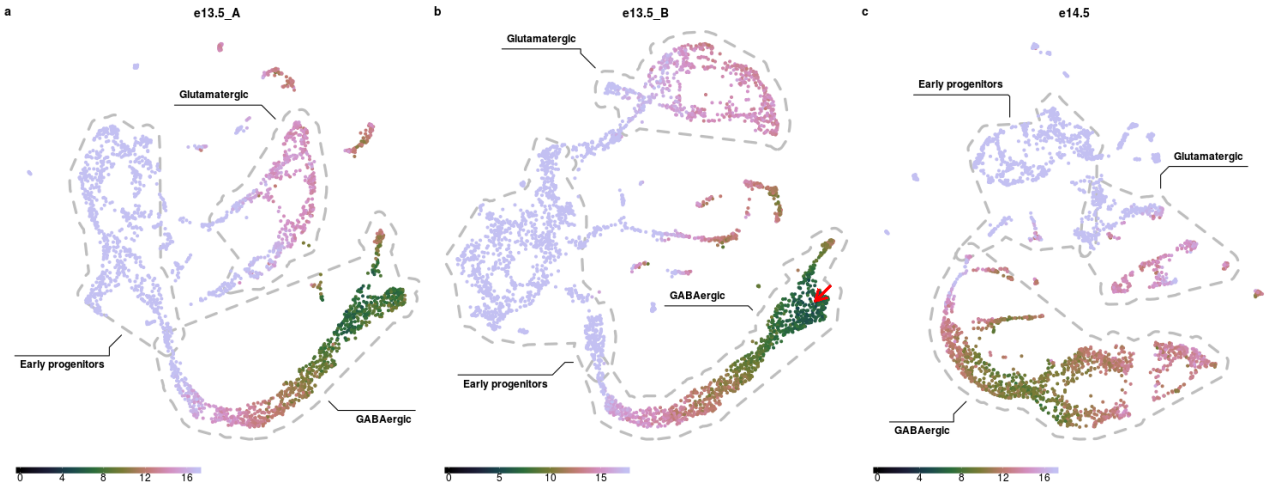



FIG. 4: Sleepwalk in multi-sample mode, comparing three samples of a developing murine cerebellum: two samples of two different mice embryos at time point E13.5 (a, b) and the third (c) from E14.5. The red arrow shows the current mouse position. The dashed grey lines roughly indicates two different lineages and their common progenitor cells. By following the GABAergic branch one can notice that its very tip in the E13.5 samples corresponds to cells in the middle of the branch in E14.5, indicating that the cells have differentiated further during the elapsed day.

For such comparisons, Sleepwalk offers a variant to the mode for embedding comparisons described above, in which points that correspond to the same cell will get different colours in different panels of the app, each showing the same embedding but having a different distance matrix assigned to it. By hovering the mouse over a cell, the user can see how the cells neighbourhoods differ change between the distance metrics.

In the life HTML supplement, we demonstrate this using the cord-blood dataset.

### G. Beyond single-cell transcriptomics

In all the examples discussed so far, the points correspond to individual cells in samples assessed with single-cell transcriptomics. Another important use case for dimension-reduction methods are large-scale studies comprising dozens or even hundreds of samples. Ref. [20], for example, describes a collection of 131 bulk RNA-Seq data sets comparing organ samples from several species and provide an overview PCA plot as their Figure 1, Ref. [21] uses a t-SNE plot to illustrate similarities and differences between their 246 blood cancer samples. Clearly, Sleepwalk can also be useful to explore dimension-reduced embeddings arising in such applications.

Research on dimension reduction originated in the machine learning field, with the orginal applications being the study of training data for machine learning applications. Of course, in this area, as well as in other applications of dimension reduction, Sleepwalk should also prove useful.

### H. Usage

Sleepwalk is provided as a package for the statistical programming language R. The main function of the package is also called "sleepwalk". The user provides it with the 2D coordinates for each object (cell) in the embedding and a square matrix of cell-to-cell distances, or, alternative to the latter, a data matrix from which sleepwalk can calculate Euclidean distances. For both these parameters, the user can also supply multiple matrices in order to display multiple embeddings concurrently for comparison. This can be done either such that each embedding represents the same objects (as in Figure 3), or that each embedding represents a different set of objects but distances are given also between objects in different embeddings (as in Figure 4).

Sleepwalk can easily be used in combination with other single-cell analysis frameworks. Visualizing, for example, a Seurat data object can be done with one line of code, as explained on the documentation web page.

By default, the sleepwalk function displays the visualization app in a web browser. Alternatively, it can also write it into an HTML file, which can then be opened with any web browser without the need of having R or the sleepwalk package running or even installed. This is useful when an analyst wishes to share a sleepwalk visualization with colleagues or provide it on a web page or in a paper supplement.

For a description of further options of the function, please see the documentation.

The app offers a "lasso" functionality: The user can encircle a group of points with the mouse, and the indices of these points are then reported back to the R session, where they can be queried with a package function. This is helpful if the analyst spots an interesting set of cells while exploring an embedding and wishes to perform further analyses on them.

We also mention the "slw_snapshot" function, which produces static plots, like the figures in this paper.

## III. DISCUSSION AND CONCLUSION

Dimension-reduced embedding such as those provided by t-SNE and UMAP have become a core tool in single-cell transcriptomics. They provide an overview of a study, help to check for expected and unexpected features in the data, allow researchers to form new hypotheses and to plan and organise the subsequent analysis. As they unavoidably contain artefacts, a common concern is that these plots may be over-interpreted.

Dimension reduction is a research area with a rich history, long predating the use of these techniques for single-cell biology. The issue with distortions has been long discussed, with the possible distortions being classified (e.g., [22]) and quantified (e.g., [23]), and advice on careful interpretation derived from these (e.g., [24]). To visually alert the viewer to distortions, some authors have suggested to colour each point by its so-called stress, i.e., the deviation of the point's on-screen distance to the other points from the distances in feature space [25] and others to colour the area around the points according to the amount of compression or stretching that the manifold underwent locally due to projection [26].

Such visualizations are useful tools for developing and improving dimension reducing methods. Our approach, however, offers a novel aspect that is crucial: rather than merely alerting the user to distortions, Sleepwalk allows the user to directly see the underlying "truth" for the selected cell. This is possible due to our use of interactivity: by allowing the user to rapidly move focus from cell to cell, and the app instantly following in redrawing the colours, we are effectively escaping the confines of a two-dimensional representation (or, three-dimensional, if we also count static colouring as a dimension).

We have shown how this novel approach gives insights into dimension-reduced embeddings that would otherwise stay hidden and thus solves a core problem in the practical use of dimension reduction methods. We envision that Sleepwalk will be used in two manners: first as a tool of exploratory data analysis, helping researchers to better understand their data, but also second as a reporting and communication tool, allowing researchers to present their results in a more transparent way. For this latter application, Sleepwalk's ability to produce stand-alone HTML pages is crucial, as these pages can then be used, e.g., as supplements to publications, where they allow readers to check embeddings themselves, without the need to install any software.

We should be clear that a visual, interactive data exploration with Sleepwalk does not replace formal

inference but complements or typically precedes it. Once one has formed a hypothesis about one's data using Sleepwalk, one should employ suitable formal analysis methods, such as statistical hypotheses tests, to confirm them. That analysis will then typically be done on the full, high-dimensional data. Dimension reduction methods are data reduction methods: this sacrifice of data is done to allow for visual inspection, but is a hindrance for any numerical analysis.

The principle of Sleepwalk's interactive is useful not only for inspection of a single data set but also lends itself for generalization to comparative tasks. We have shown several possible modes of comparison: between different embeddings of the same data, between embeddings from several samples, and between different ways of preprocessing the data and obtaining distances. The comparison between samples will find direct application in any study working with multiple samples, the other two are useful in method selection and in method development, as they allow for comparison of data processing pipelines.

We therefore expect that Sleepwalk will find broad use not only in single-cell transcriptomics, but essentially all instances of big data where experimental units (cells, samples, or the like) are descibed in a high-dimensional feature space.

### Availability

The "sleepwalk" R package is available on the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/package=sleepwalk, released as open-source software under the GNU General Public License v3 (or later). Documentation, installation instructions and examples for Sleepwalk can be found on the project web page at https://anders-biostat.github.io/sleepwalk/. The source code is available on GitHub (https://github.com/anders-biostat/sleepwalk),

### Acknowledgments

We would like to thank Kevin Leiss for helpful discussions and assistance in preprocessing the cerebellum data with CellRanger. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) via the collaborative research centres SFB 1036 (S.A.) and SFB 1366 (S.O.).

## IV. METHODS

### A. Sleepwalk implementation

Sleepwalk is written in JavaScript, using the D3.js data visualization framework [27]. JavaScript was chosen because it is available on all common platforms, usually without need to install anything, thus enabling the standalone HTML feature, because writing the app with D3.js was convenient, and because the JavaScript engines of most web browsers offer very good performance, enabling smooth rendering of the colour changes and thus the instant interactive feedback that is required to provide an intuitive user experience.

A thin wrapper of R code around the JavaScript code turns Sleepwalk into an R package, allowing for convenient integration into currently popular workflows for single-cell analysis like Seurat. We use the httpuv R package [28] to bridge between the R session and the web browser. It sets up a simple local server to serve the app to the browser and then use its implementation of the WebSocket protocol [29] to keep open a communication channel between R session and web browser. This allows the app to report back to the R session when the user has selected points using the lasso feature or to make snapshots and change the state of the app from an R session.

The color scheme used to depict distances is the "cubehelix" palette, a colour map originally developed for astronomy and optimized for good visual separation between levels throughout its dynamic range [30].

### B. Example data

#### 1. *Cord-blood data set*

The cord-blood data from [4] are available at Gene Expression Omnibus (GEO) via accession GSE100866. The raw UMI counts were processed following the Seurat workflow proposed for exactly this data set [6]. Data were normalized and log transformed. 976 variable genes were detected with y.cutoff = 0.5. These genes were scaled and used for principle components analysis. For further analysis, the first 13 principal components were used, which explain around 23% of the total variance. The t-SNE (Figures 1, 2, supplement files) and UMAP (supplement files) embeddings were calculated using the default functions from the Seurat package. The assignments of cell types to clusters was taken, too, from the Seurat tutorial workflow [6]. Normalized and log-transformed epitome data were used to calculate Euclidean distances for the distance comparison (see supplement files). The resulting Seurat object can be downloaded from figshare.com (doi:10.6084/m9.figshare.7908059).

### 2. Murine cerebellum data set

The raw sequence data from [16] were downloaded from the European Nucleotide Archive under the accession number PRJEB23051. The reads were aligned and counted using the Cell Ranger [31] software (output files are accessible from figshare.com, doi:10.6084/m9.figshare.7910483). Some genes and droplets were filtered out following the Methods section of [16]. We removed all the cells with more than 10% of all UMIs coming from mitochondrial genes. We then removed all ribosomal and mitochondrial genes. Finally, only cells that contain from 3500 to 15000 UMIs were kept. Lastly, we omitted all genes with zero expression in all the cells. The filtered raw data were then used to create Seurat objects that can be found at figshare.com, doi:10.6084/m9.figshare.7910483), or on GitHub https://github.com/anders-biostat/sleepwalk/tree/paper in the "data" folder. We used Seurat to normalise and log-transform raw counts and find variable genes. The irlba R package [32, 33] was used to generate a PCA embedding of the data (each sample separately, only variable genes). The first 50 principal components were used for further analysis. To render a t-SNE embedding we used the Rtsne package [34]; the uwot package [35] was used for UMAP embeddings. To calculate distances between cells from different samples (Figure 4), we used the variable genes shared between all the samples and produces a PCA embedding for all the cells. Euclidean distances in the space defined by the first 50 principal components are used to colour the points. To distinguish early progenitors form further differentiated cells of glutamatergic and GABAergic we used the following marker genes: Msx3 for early progenitors, Meis2 for the glutamatergic lineage, adn Lhx5 for the GABAergic lineage (Supplementary Figure 1). Countours in Figure 4 are drawn to include around 90% of cells that express each of the markers above a certain threshold using the "geom_mark_hull" function of the ggforce package [36].

The R scripts used to produce all the figures and supplement materials are available at https://github.com/anders-biostat/sleepwalk/tree/paper.


[1] Ringnér, M. *What is principal component analysis?* Nature Biotechnology 26:303 (2008). [Link].

[2] Kruskal, J.B. *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.* Psychometrika 29:1 (1964). [Link].

[3] Kohonen, T. *Self-organized formation of topologically correct feature maps.* Biological Cybernetics 43:59 (1982). [Link].

[4] Stoeckius, M., et al. *Simultaneous epitope and transcriptome measurement in single cells.* Nature Methods 14:865 (2017). [Link].

[5] Butler, A., et al. *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nature Biotechnology 36:411 (2018). [Link].

[6] Satija Lab. *Using Seurat with multi-modal data* (2018). URL https://satijalab.org/seurat/multimodal_vignette.html.

[7] van der Maaten, L.J.P., et al. *Visualizing high-dimensional data using t-SNE.* Journal of Machine Learning Research 9:2579 (2008). [Link].

[8] McInnes, L., et al. *UMAP: Uniform manifold approximation and projection for dimension reduction.* arXiv:1802.03426 [cs, stat] (2018). URL http://arxiv.org/abs/1802.03426.

[9] Coifman, R.R., et al. *Diffusion maps.* Applied and Computational Harmonic Analysis 21:5 (2006).

[10] Angerer, P., et al. *destiny: diffusion maps for large-scale single-cell data in R.* Bioinformatics 32:1241 (2016). [Link].

[11] Trapnell, C., et al. *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nature Biotechnology 32:381 (2014). [Link].

[12] Qiu, X., et al. *Reversed graph embedding resolves complex single-cell trajectories.* Nature Methods 14:979 (2017). [Link].

[13] Mao, Q., et al. *Dimensionality reduction via graph structure learning.* In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 765–774. ACM, New York, NY, USA (2015). [Link].

[14] Waltman, L., et al. *A smart local moving algorithm for large-scale modularity-based community detection.* The European Physical Journal B 86:471 (2013). [Link].

[15] Becht, E., et al. *Dimensionality reduction for visualizing single-cell data using UMAP.* Nature Biotechnology 37:38 (2019). [Link].

[16] Carter, R.A., et al. *A single-cell transcriptional atlas of the developing murine cerebellum.* Current Biology 28:2910 (2018). [Link].

[17] Becker, R.A., et al. *Brushing scatterplots.* Technometrics 29:127 (1987). [Link].

[18] Tung, P.Y., et al. *Batch effects and the effective design of single-cell gene expression studies.* Scientific Reports 7:39921 (2017). [Link].

[19] Haghverdi, L., et al. *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.* Nature Biotechnology 36:421 (2018). [Link].

[20] Brawand, D., et al. *The evolution of gene expression levels in mammalian organs.* Nature 478:343 (2011). [Link].

[21] Dietrich, S., et al. *Drug-perturbation-based stratification of blood cancer.* Journal of Clinical Investigation 128:427 (2018). [Link].

[22] Aupetit, M. *Visualizing distortions and recovering topology in continuous projection techniques.* Neurocomputing 70:1304 (2007). [Link].

[23] Kaski, S., et al. *Trustworthiness and metrics in visualizing similarity of gene expression.* BMC Bioinformatics 4:48 (2003). [Link].

[24] Wattenberg, M., et al. *How to use t-SNE effectively.*

Distill (2016). doi:10.23915/distill.00002.

[25] Seifert, C., et al. *Stress maps: Analysing local phenomena in dimensionality reduction based visualisations*. In: Kohlmahher, J., et al. (Editors) *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE (2010).

[26] Lespinats, S., et al. *CheckViz: Sanity check and topological clues for linear and non-linear mappings*. Computer Graphics Forum 30:113 (2011). [Link].

[27] Bostock, M., et al. *$D^3$ data-driven documents*. IEEE Transactions on Visualization and Computer Graphics 17:2301 (2011). [Link].

[28] Cheng, J., et al. *httpuv: HTTP and WebSocket server library* (2019). URL https://CRAN.R-project.org/package=httpuv. R package version 1.5.0.

[29] Fette, I., et al. *RFC 6455: The WebSocket protocol*. Internet Engineering Task Force (2011). [Link].

[30] Green, D.A. *A colour scheme for the display of astronomical intensity images*. Bulletin of the Astromical Society of India 39:289 (2011).

[31] 10x Genomics. *What is Cell Ranger?* URL https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger.

[32] Baglama, J., et al. *irlba: Fast truncated singular value decomposition and principal components analysis for large dense and sparse matrices* (2019). URL https://CRAN.R-project.org/package=irlba. R package version 2.3.3.

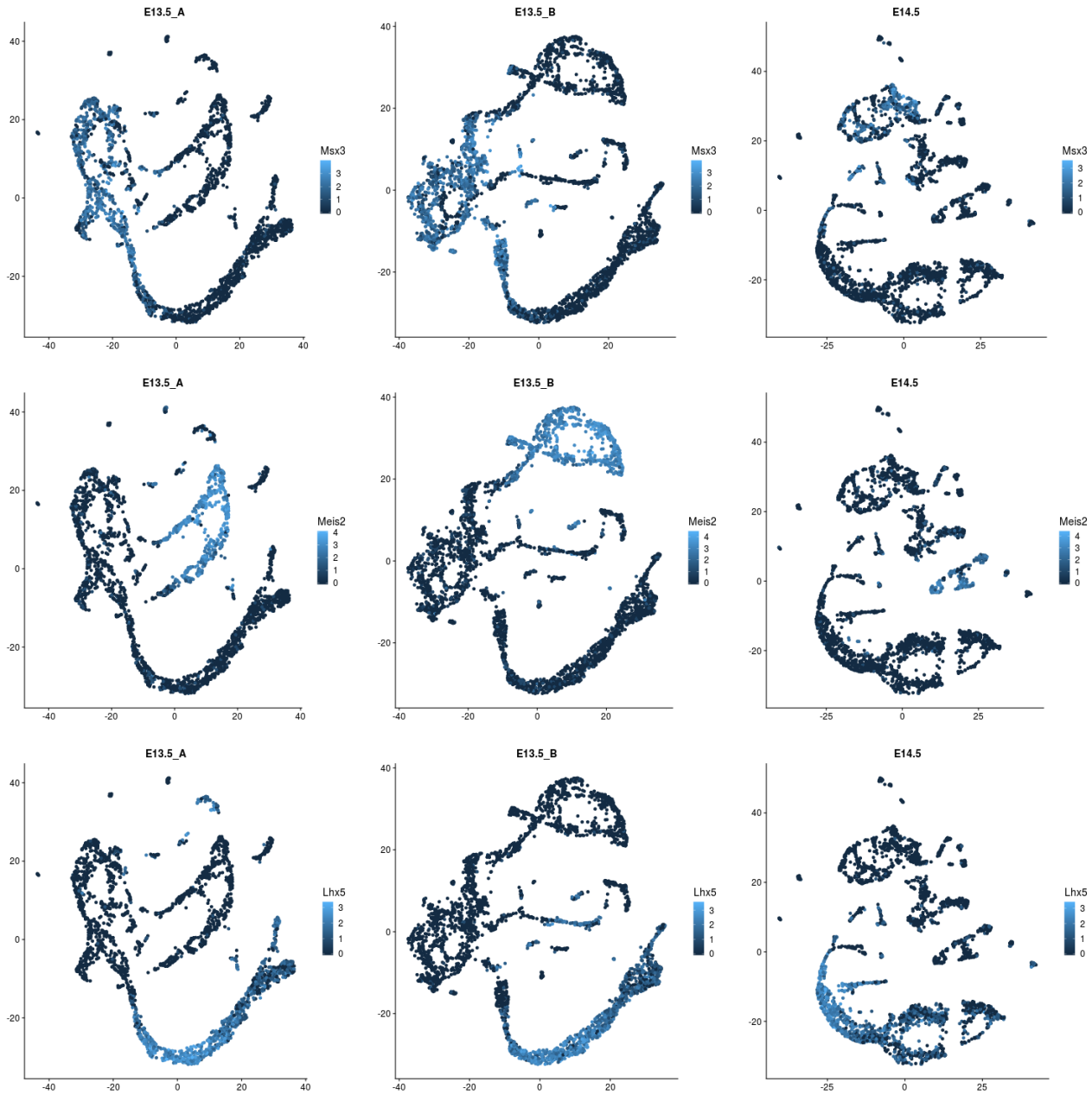
[33] Baglama, J., et al. *Augmented implicitly restarted Lanczos bidiagonalization methods*. SIAM Journal on Scientific Computing 27:19 (2005). [Link].

[34] Krijthe, J.H. *Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation* (2015). URL https://github.com/jkrijthe/Rtsne. R package version 0.15.

[35] Melville, J. *uwot: The uniform manifold approximation and projection (UMAP) method for dimensionality reduction* (2019). URL https://github.com/jlmelville/uwot. R package version 0.0.0.9010.

[36] Pedersen, T.L. *ggforce: Accelerating ggplot2* (2019). URL https://CRAN.R-project.org/package=ggforce. R package version 0.2.0.

## Appendix A: Supplementary Figure




SUPPL. FIG. 1: Markers used to identify lineages in Figure 4.

## Appendix B: HTML Supplement

This supplement is provided as an enclosed HTML file. Alternatively, it can be found online at https://anders-biostat.github.io/sleepwalk/supplementary/.