[Skip to main content](https://www.sciencedirect.com/science/article/pii/S2405471218302357#app)

[Journals](https://www.sciencedirect.com/science/journals)[Books](https://www.sciencedirect.com/science/bookbshsrw)

My account



[Download PDF](https://www.sciencedirect.com/science/article/pii/S2405471218302357/pdfft?md5=520fc56d074a2e9d5e17fb9d00f4db44&pid=1-s2.0-S2405471218302357-main.pdf)

Export

Top of Form

[Advanced](https://www.sciencedirect.com/search/advanced)

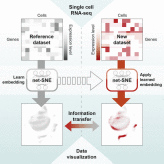
Bottom of Form

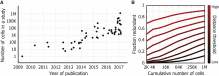
Outline

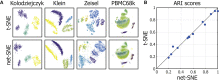
1. [Highlights](https://www.sciencedirect.com/science/article/pii/S2405471218302357#abs0020)
2. [Summary](https://www.sciencedirect.com/science/article/pii/S2405471218302357#abs0010)
3. [Graphical Abstract](https://www.sciencedirect.com/science/article/pii/S2405471218302357#abs0015)
4. [Keywords](https://www.sciencedirect.com/science/article/pii/S2405471218302357#kwrds0010)
5. [Introduction](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec1)
6. [Results](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec2)
7. [Discussion](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec3)
8. [STAR★Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)
9. [Acknowledgments](https://www.sciencedirect.com/science/article/pii/S2405471218302357#ack0010)
10. [Supplemental Information](https://www.sciencedirect.com/science/article/pii/S2405471218302357#appsec2)
11. [References](https://www.sciencedirect.com/science/article/pii/S2405471218302357#cebib0010)

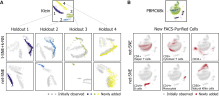
Show full outline

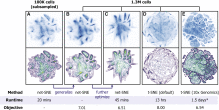
Figures (5)

[[](https://www.sciencedirect.com/science/article/pii/S2405471218302357#undfig1)](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "undfig1)

[[](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig1)](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig1)

[[](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig2)](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig2)

[[](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig3)](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig3)

[[](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig4)](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig4)

Tables (1)

1. [Table](https://www.sciencedirect.com/science/article/pii/S2405471218302357#undtbl1)

Extras (2)

1. [Document S1. Figures S1–S4 and Table S1](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)
2. [Document S2. Article plus Supplemental Information](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc2)

Available online 20 June 2018

[In Press, Corrected Proof](https://www.sciencedirect.com/science/journal/aip/24054712)[What are Corrected Proof articles?](https://service.elsevier.com/app/answers/detail/a_id/22801/supporthub/sciencedirect/)

[](https://www.sciencedirect.com/science/journal/24054712)

Report

Generalizable and Scalable Visualization of Single-Cell Data Using Neural Networks

Author links open overlay panel[HyunghoonCho](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "!)[1](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "!)[BonnieBerger](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "!)[124](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "!)[JianPeng](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "!)[3](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "!)

1

Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

2

Department of Mathematics, MIT, Cambridge, MA 02139, USA

3

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received 9 March 2018, Revised 12 April 2018, Accepted 12 May 2018, Available online 20 June 2018.

Published: June 20, 2018

Show less

<https://doi.org/10.1016/j.cels.2018.05.017>[Get rights and content](https://s100.copyright.com/AppDispatchServlet?publisherName=ELS&contentID=S2405471218302357&orderBeanReset=true)

Highlights

•

We train a neural network to visualize single-cell RNA-sequencing datasets

•

Our method can map new cells onto existing visualizations to allow knowledge transfer

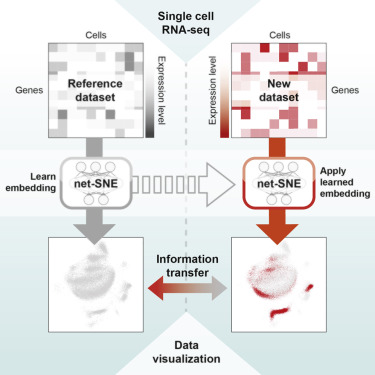
•

Our method efficiently visualizes millions of cells via a bootstrap procedure

Summary

Visualization algorithms are fundamental tools for interpreting single-cell data. However, standard methods, such as t-stochastic neighbor embedding (t-SNE), are not scalable to datasets with millions of cells and the resulting visualizations cannot be generalized to analyze new datasets. Here we introduce net-SNE, a generalizable visualization approach that trains a neural network to learn a mapping function from high-dimensional single-cell gene-expression profiles to a low-dimensional visualization. We benchmark net-SNE on 13 different datasets, and show that it achieves visualization quality and clustering accuracy comparable with t-SNE. Additionally we show that the mapping function learned by net-SNE can accurately position entire new subtypes of cells from previously unseen datasets and can also be used to reduce the runtime of visualizing 1.3 million cells by 36-fold (from 1.5 days to an hour). Our work provides a framework for bootstrapping [single-cell analysis](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/single-cell-analysis)from existing datasets.

Graphical Abstract



1. [Download high-res image (157KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-fx1_lrg.jpg)
2. [Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-fx1.jpg)

Keywords

data visualization

single-cell RNA sequencing

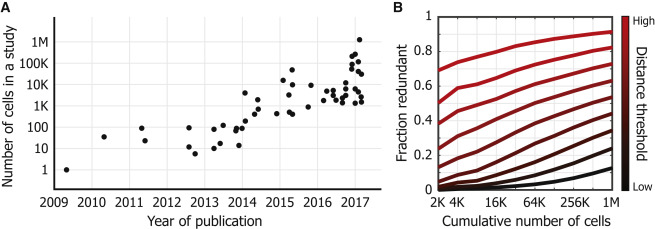
neural network

Introduction

Complex biological systems arise from functionally diverse, heterogeneous populations of cells. Single-cell [RNA sequencing](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing) (scRNA-seq) ([Gawad et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib9)), which profiles [transcriptomes](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/transcriptome) of individual cells rather than bulk samples, has been a key tool in dissecting the intercellular variation in a wide range of domains, including cancer biology ([Wang et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib45)), immunology ([Stubbington et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib35)), and [metagenomics](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/metagenomics) ([Yoon et al., 2011](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib46)). scRNA-seq also enables the *de novo* identification of cell types with distinct expression patterns ([Grün et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib11), [Jaitin et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib16)).

A standard analysis for scRNA-seq data is to visualize single-cell gene-expression patterns of samples in a low-dimensional (2D or 3D) space via methods such as t-stochastic neighbor embedding (t-SNE) ([Maaten and Hinton, 2008](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib23)) or, in earlier studies, principal component analysis ([Jackson, 2005](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib15)), whereby each cell is represented as a dot and cells with similar [expression profiles](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gene-expression-profiling) are located close to each other. Such visualization reveals the salient structure of the data in a form that is easy for researchers to grasp and further manipulate. For instance, researchers can quickly identify distinct subpopulations of cells through visual inspection of the image, or use the image as a common lens through which different aspects of the cells are compared. The latter is typically achieved by overlaying additional data on top of the visualization, such as known labels of the cells or the expression levels of a gene of interest ([Zheng et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib49)). While many of these approaches have initially been explored for visualizing bulk [RNA-seq](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/rna-seq) ([Palmer et al., 2012](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib25), [Simmons et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib34)), methods that take into account the idiosyncrasies of scRNA-seq (e.g., dropout events where nonzero expression levels are missed as zero) have also been proposed ([Pierson and Yau, 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib28), [Wang et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib44)). Recently, more advanced approaches that visualize the cells while capturing important global structures such as cellular hierarchy or trajectory have been proposed ([Anchang et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib3), [Hutchison et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib14), [Moon et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib24), [Qiu et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib30)), which constitute a valuable complementary approach to general-purpose methods such as t-SNE.

Comprehensively characterizing the landscape of single cells requires a large number of cells to be [sequenced](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing). Fortunately, advances in automatic cell isolation and multiplex sequencing have led to an exponential growth in the number of cells sequenced for individual studies ([Svensson et al., 2018](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib36)) ([Figure 1](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig1)A). For example, 10x Genomics recently made publicly available a dataset containing the expression profiles of 1.3 million brain cells from mice (<https://support.10xgenomics.com/single-cell-gene-expression/datasets>). However, the emergence of such mega-scale datasets poses new computational challenges before they can be widely adopted. Many of the existing computational methods for analyzing scRNA-seq data require prohibitive runtimes or computational resources; in particular, the state-of-the-art implementation of t-SNE ([Van Der Maaten, 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib42)) requires 1.5 days to run on 1.3 million cells based on our estimates.



1. [Download high-res image (375KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr1_lrg.jpg)
2. [Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr1.jpg)

Figure 1. The Increasing Scale and Redundancy of Single-Cell [RNA-Seq](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/rna-seq) Datasets

(A) The exponential increase in the number of single cells [sequenced](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing) by individual studies (adapted from [Svensson et al., 2018](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib36)). Note that the y axis scales exponentially.

(B) Retrospective analysis of redundancy in the Brain1m dataset ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "sec4)) with 2,000 initial cells and repeated doubling of the data size. For each batch added, we computed the distribution of the cells' minimum Euclidean distance to cells already observed based on their gene expression. Each curve corresponds to a particular distance threshold for deeming the new cell redundant. The thresholds are chosen as the deciles of the overall distribution of minimum Euclidean distances.

Here, we introduce neural t-SNE (net-SNE), a scalable and generalizable method for visualizing single cells for scRNA-seq analysis. Taking inspiration from compressive genomics ([Loh et al., 2012](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib22)), we exploit the intuition that when a large number of cells are sequenced, a significant portion of the cells are redundant (i.e., highly similar to other cells). Taking advantage of the expressive power of neural networks (NNs), which has been demonstrated in numerous applications ([LeCun et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib21)), net-SNE trains an NN to learn a high-quality mapping function that takes an expression profile as input and outputs a low-dimensional embedding in 2D or 3D for visualization. Unlike t-SNE, the mapping function learned by net-SNE can be used to map previously unseen cells that were not included in the input data. This capability allows for novel workflows for single-cell studies, whereby newly observed cells are visualized in the context of existing datasets to gain additional insights.

To demonstrate visualization quality as well as scalability, we show that net-SNE learns visualizations that are similar to those of t-SNE on 14 scRNA-seq datasets of various cell types and data sizes up to 1.3 million cells. Next we focused on generalizability, demonstrating that net-SNE newly achieves the ability to map previously unseen cells; in particular, we show that net-SNE not only can identify subtypes of cells that were not included in the initial data but can also be used to bootstrap the visualization from a subset of data to achieve significantly better scalability to mega-scale datasets. Given the inherent redundancy in biological data ([Yu et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib47)), we expect our techniques for neural data visualization to accelerate and enhance other high-dimensional biological data analyses beyond visualization.

Results

Increasing Redundancy in Single-Cell Datasets

We first set out to empirically assess the extent to which additional [sequencing](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing) of single cells from the same biological source capture unforeseen expression patterns. Starting with 2,000 randomly chosen cells from the 10x Genomics scRNA-seq dataset with 1.3 million mouse neurons ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)), we repeatedly doubled the data size up to a million cells by sampling the remaining cells (without replacement) and measured how redundant the newly added cells are compared with the ones already observed ([Figure 1](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig1)B). As the scale of data grows, sequencing more cells exhibits a clear diminishing return in terms of capturing cells with unique expression patterns. For example, the majority of cells (53%) in the final half of the data can be considered redundant according to a certain distance threshold, which deems only 10% of the cells redundant in the initial batch. Nevertheless, a considerable fraction of the newly observed cells remains unique even at the scale of a million cells. For instance, 10% of the cells in the final half are as unique as the top 30% of the initial batch, which suggests that the push toward a higher cell count is indeed valuable for gaining access to relatively unexplored regions of the gene-expression landscape, albeit with decreasing effectiveness.

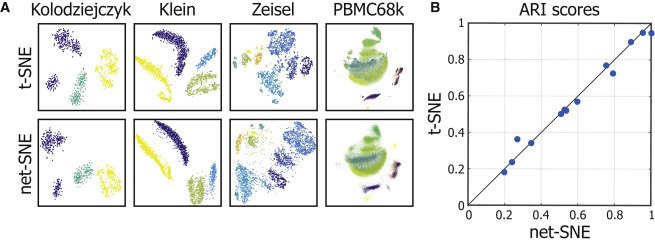
These results imply that, as researchers collectively accumulate scRNA-seq data for a particular biological system (e.g., tissue, organism, or microbial population) to the scale of millions of cells, a significant portion of newly sequenced cells will fall into the space already visited by existing data, where useful insights may be available from previous analyses. This observation motivates our development of net-SNE, which allows new data to be mapped onto an existing visualization to accelerate such knowledge transfer across different studies or experiments.

Overview of net-SNE

net-SNE achieves generalizability by training a feedforward neural network ([LeCun et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib21)) to learn a parameterized embedding function that takes a cell's [expression profile](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gene-expression-profiling) as input and outputs the coordinates in a low-dimensional space for visualization ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)). Given the wide success of t-SNE in single-cell biology ([Amir et al., 2013](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib1)), we aim to emulate the behavior of t-SNE while newly achieving the ability to map new cells, by training our neural network to optimize the same objective function as t-SNE. This objective function intuitively captures how faithfully the local structure among the input vectors (i.e., single-cell gene-expression profiles) is represented in the visualization. Although our *parametric* approach to t-SNE has been theoretically considered ([Van Der Maaten, 2009](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib43)), its application to real-world, large-scale datasets has been considerably limited due to the difficulties in successfully training a neural network to perform a complex task such as t-SNE. Our work employs new optimization techniques to improve the scalability of neural network training for t-SNE ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)) and newly demonstrates the effectiveness of this approach for [single-cell analysis](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/single-cell-analysis).

net-SNE Learns High-Quality Visualizations of Single Cells

To evaluate the ability of net-SNE to accurately model the visualization of single-cell datasets, we tested it on 13 existing scRNA-seq datasets of varying sizes *with known clusters* ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)). We found that for all of the datasets net-SNE is able to learn an embedding that closely matches the output of t-SNE ([Figures 2](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig2)A and [S1](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "mmc1)).



1. [Download high-res image (577KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr2_lrg.jpg)
2. [Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr2.jpg)

Figure 2. net-SNE Recapitulates t-SNE Mapping on 13 Benchmark Datasets with Known Subtypes

(A) Comparison of net-SNE and t-SNE visualizations on four largest benchmark datasets with known clusters. Colors indicate known cell types provided by the original work. net-SNE visualizations of the Klein and Zeisel datasets are reflected over a diagonal axis for comparison. Figures for the remaining datasets are provided in [Figure S1](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1).

(B) Quality of each visualization is quantified by the adjusted Rand index between the known labels and the output of k-means clustering based on the embedding. Each dot represents one of the 13 datasets analyzed. Results based on agglomerative clustering instead of k-means are provided in [Figure S2](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1), also showing a high concordance between net-SNE and t-SNE.

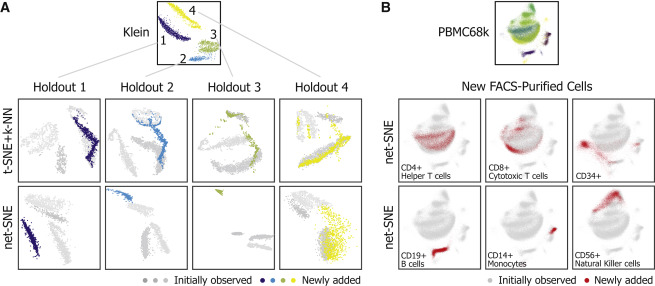
To systematically evaluate the quality of embeddings produced by net-SNE, we assessed the agreement between the known subtypes and clusters that are computationally identified based on the low-dimensional embeddings. The level of agreement was quantified by the adjusted Rand index ([Rand, 1971](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib31)), following previous work ([Kiselev et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib18)). We obtained the clusters by applying the standard k-means clustering algorithm ([Hartigan and Wong, 1979](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib13)) to the embeddings, where the known number of clusters was provided to the algorithm. As shown in [Figure 2](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig2)B, net-SNE achieves clustering accuracy that is comparable with t-SNE for all 13 datasets, which agrees with the visual concordance of the two methods. An analogous analysis we performed, based on agglomerative clustering instead of k-means, leads to similar concordance between net-SNE and t-SNE ([Figure S2](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)).

Notably, we obtained all of these results using a relatively simple neural network with only two layers of nonlinearities with 50 units in each layer ([Figure S1](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)B). In additional experiments, not only did we observe that the net-SNE results are reasonably stable across a wide range of network architectures, we also found that the size of our network can be reduced to as low as 10 units per layer without significantly sacrificing the quality of visualizations ([Figure S4](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)), even for the PBMC68k dataset containing tens of thousands of cells. This finding suggests that the relationship between gene-expression profiles and the clustering pattern of cells may be simple enough to admit a concise characterization for various cell populations.

net-SNE Accurately Maps New Cells

To demonstrate the potential of net-SNE for translational analyses across different datasets, we performed a cross-validation experiment whereby an entire cluster of cells was removed from a dataset and placed onto the visualization after the fact. While the original t-SNE does not support the visualization of new data points, we considered as baseline a naive extension of t-SNE whereby the embedding of a new cell is determined as the average position (in the low-dimensional space) of the cell's nearest neighbors in the initial data according to expression measurements (t-SNE + k-NN). An alternative extension of t-SNE whereby the new cells are randomly initialized and optimized while fixing the positions of the initial cells similarly lacks the scalability to mega-scale datasets as the original t-SNE and thus was not considered in our analysis.

[Figure 3](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig3) shows our cross-validation results from the Klein dataset ([Klein et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib19)), which contains four known clusters, each of which was held out in four separate experiments. Remarkably, in three out of the four cases, the embedding learned by net-SNE accurately positioned the held-out cells as a distinct cluster, despite the fact that the training data did not contain any of the cells from this cluster. In contrast, our nearest neighbor-extension of t-SNE (t-SNE + k-NN) overlaid most of the new cells onto existing clusters and ended up incorrectly outputting an obfuscated map. Although visualizing the entire dataset from scratch tends to result in better-quality scores than both of these approaches ([Figure S4](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)A), we note that the initial generalization obtained by either approach can be further optimized if desired.



1. [Download high-res image (607KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr3_lrg.jpg)
2. [Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr3.jpg)

Figure 3. net-SNE Generalizes to Unseen Cells

(A) Each column represents a cross-validation experiment whereby the visualization is performed only on a subset of data (shown in gray), where all cells corresponding to one of the four known cell types in the Klein dataset were held out. Held-out cells were later added to the visualization (shown in color). We compared net-SNE with a naive extension of t-SNE whereby each new cell is placed at the average position of the five nearest neighbors in the initial dataset. Unlike t-SNE, net-SNE is able to identify the newly added cells as a separate cluster. Also, in the setting where the new cells belong to subtypes already represented in the initial dataset, net-SNE accurately assigns the new cells to the respective clusters ([Figure S4](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)B).

(B) To further test net-SNE's generalizability, we obtained scRNA-seq datasets from 10x Genomics for six representative blood cell subtypes (shown in subfigures), each purified via [fluorescence-activated cell sorting](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/flow-cytometry). We visualized each of the purified cell populations (in red) over a pre-trained net-SNE embedding of the PBMC68k dataset (shown in gray), which is from a whole blood sample encompassing all of the purified cell types. Despite coming from a different [sequencing](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing) experiment with different sample preparation, many of the purified cell types were mapped by net-SNE to well-defined clusters in the visualization, demonstrating its potential for knowledge transfer across different datasets.

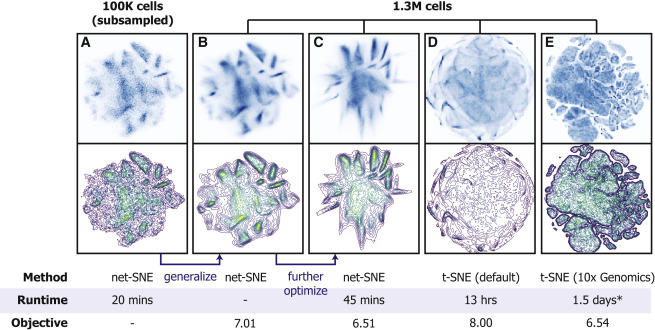
The setting of this cross-validation analysis may arise in practice in cases where a rare subpopulation of cells was omitted from the initial dataset (e.g., due to small data size). Our results suggest that, while the naive nearest neighbor-based projection of newly observed cells (including the rare subtype) will likely render the new subtype invisible, net-SNE is still able to identify the new cluster, given that its gene expression is sufficiently distinct from that of existing cells. In an alternative setting where the new cells are from a subtype that is already represented in the initial dataset, net-SNE is still able to accurately assign the new cells to the correct cluster ([Figure S4](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)B).

To further demonstrate the utility of net-SNE's generalization performance beyond cross-validation, we obtained six scRNA-seq datasets of different purified blood cell subtypes from 10x Genomics ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)). We projected each dataset onto a pre-trained, net-SNE visualization of the PBMC68k dataset of a whole blood sample, which includes all six subtypes. Despite the differences in sample preparation and the possibility of batch effects ([Tung et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib40)), net-SNE accurately positions the purified cell populations onto existing clusters, immediately providing a useful characterization for a number of clusters in the PBMC68k dataset ([Figure 3](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig3)B). Notably, net-SNE projected most of the purified CD34-positive cells onto a distinct region that contained only a small number of cells in the initial dataset. This observation is consistent with the low basal levels of CD34-positive cells in blood ([Kikuchi-Taura et al., 2006](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib17)) and further illustrates net-SNE's ability to capture previously unseen subtypes of cells.

net-SNE Accelerates Visualization of Millions of Cells

After validating net-SNE's ability to map new cells, we then asked whether this ability can be exploited to achieve fast visualization of mega-scale datasets. Drawing from the intuition that datasets of this scale can be accurately represented by a smaller subset of cells due to high redundancy, we first trained net-SNE on a subset of 100,000 cells from the Brain1m dataset containing 1.3 million cells and later applied the learned embedding function to the entire dataset. This fast approach took around only 20 min overall and resulted in a higher-quality map than the output of t-SNE with the default parameter settings, which took 13 hr to finish. Note that we use the Kullback-Leibler (KL) divergence objective score—the quantity minimized by both t-SNE and net-SNE—as the metric of quality (inversely related), which is more objective than a visual assessment. If a researcher already has access to a pre-trained mapping based on an existing dataset, the reduction in runtime achieved by net-SNE is likely to be even more drastic (e.g., days for t-SNE to a few minutes).

It is worth noting that the original dataset provided by 10x Genomics also included a t-SNE embedding, which appeared higher in quality than the t-SNE output we obtained using the default setting ([Figure 4](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "fig4)E). While the objective score we computed based on the published embedding was superior to net-SNE's initial generalization, 45 min of further optimization of net-SNE was sufficient to outperform this score ([Figure 4](https://www.sciencedirect.com/science/article/pii/S2405471218302357#fig4)C).



1. [Download high-res image (1MB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr4_lrg.jpg)
2. [Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-gr4.jpg)

Figure 4. net-SNE Enables Fast Visualization of Mega-Scale Datasets

The Brain1m dataset with 1.3 million cells is visualized using a novel bootstrap approach enabled by net-SNE, whereby the embedding learned on a subset of 100 K cells (A) was used to initialize training on the whole dataset. Our initial generalization to the full data instantly obtained (in minutes) a visualization (B) of higher quality, as measured by the KL divergence objective score minimized by both methods ([STAR Methods](https://www.sciencedirect.com/science/article/pii/S2405471218302357#sec4)), than that of t-SNE with default parameters achieved after 13 hr (D). While the t-SNE embedding provided in the original dataset by 10x Genomics (E) achieves a better objective, net-SNE outperforms this embedding with less than an hour of further training (C). Top row shows the heatmap for each embedding using a linear color map, where the highest value represented is chosen for each plot to achieve the best clarity. Bottom row shows contour plots of the same data. Lower objective score corresponds to better agreement between the gene-expression landscape and the visualization. ∗The visualization by 10x Genomics can be closely reproduced by increasing the number of iterations for t-SNE, and based on our experiments t-SNE required 1.5 days to achieve a solution with a comparable score.

After observing that we can achieve a visualization similar to the one provided by 10x Genomics by increasing the number of iterations for t-SNE (thus increasing runtime), we performed an experiment whereby we ran t-SNE until it reached an objective score matching the provided embedding. This resulted in a runtime estimate of 1.5 days for the published visualization, which is substantially longer than that required by net-SNE to achieve a superior quality (i.e., 20 min of pre-training and 45 min of further optimization). Notably, if the visualization by net-SNE was performed based on a well-characterized dataset, the map obtained by net-SNE would have the additional benefit of allowing researchers to immediately transfer insights from the existing dataset.

The visual difference between the net-SNE and t-SNE outputs in this experiment can be partially attributed to the fact that there are likely many locally optimal solutions to the same optimization problem solved by both methods. Although the clusters may appear more clearly separated in the t-SNE output, the fact that our visualization actually achieves a better objective score suggests the possibility of the abrupt boundaries in t-SNE being an artifact that is not warranted by the underlying data. Since net-SNE restricts the space of possible visualizations to those that can be modeled as a continuous function of gene expression as specified by our relatively simple neural network, it has a tendency to obtain a “smoother” visualization, which could potentially be a more accurate representation of the gene-expression landscape. Note that we obtained our results on the Brain1m dataset using a two-layer neural network with 50 hidden units per layer, as in our aforementioned analyses.

Discussion

As we enter the age of mega-scale [single-cell analysis](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/single-cell-analysis), new computational methods that take advantage of the growing redundancy in the scRNA-seq data are needed. To this end we have presented net-SNE, a visualization method that uses a neural network to learn a parametric embedding function that emulates t-SNE's visualization while newly achieving the ability to map previously unseen cells. We have demonstrated that net-SNE not only learns high-quality maps such as t-SNE, but also gracefully generalizes to unseen cells—even when a whole subpopulation is missing from the initial dataset or when the new data come from a different [sequencing](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sequencing) experiment. Indeed, net-SNE's ability to generalize allows researchers to exploit redundancy across different datasets by projecting the cells in one dataset onto another to facilitate transfer of knowledge. In addition, we have shown that using a pre-trained embedding from a subsampled (or an existing) dataset is an effective way for performing fast visualization of mega-scale scRNA-seq datasets. Our approach achieves significantly better scalability than t-SNE, which is on the verge of being impractical for datasets with more than a million cells. Although a number of recent studies introduced new techniques for improving the scalability of data visualization tools ([Dzwinel and Wcisło, 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib8), [Tang et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib37)), they do not address the lack of generalizability that net-SNE overcomes.

A number of recently proposed techniques for single-cell analysis can be used in conjunction with net-SNE to potentially further enhance its visualization quality; for instance, methods that account for dropout events ([Wang et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib44)) or batch effects ([Haghverdi et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib12)) can be employed to improve the input similarity matrix before applying net-SNE. Notably, [Amodio et al. (2017)](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib2), concurrently with this work, introduced an autoencoder-based approach for jointly performing batch effect correction and visualization, which finds a parametric embedding like net-SNE. Because they optimize a different objective function, however, the behavior of their embedding is fundamentally different from that of t-SNE (and thus net-SNE).

The fact that net-SNE obtains a visualization whereby the coordinates are directly modeled by the neural network as a function of gene expression opens up new directions for further research. In particular, one can investigate the parameters of the NNs trained by net-SNE for insights into what types of expression patterns are being utilized for t-SNE-like visualizations. Furthermore, while the characterization of conspicuous clusters in t-SNE output has typically been done by summarizing the expression of cells that belong to the cluster of interest, net-SNE enables a more direct and potentially more effective approach that analyzes the behavior of the embedding function.

A recently proposed idea of building a reference map of all human cell types based on high-throughput single-cell experiments (called the Human Cell Atlas [[Regev et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib32)]) is closely related to our vision. While embedding all cell types in a space with as few as two or three dimensions is unlikely to be successful given the complexity of the problem, we believe that insights from net-SNE and its future extensions may lead to an effective approach for learning compact vector space representations of all human cells that can be readily plugged into new and existing computational methods to further advance our understanding of biology.

STAR★Methods

Key Resources Table

| **REAGENT or RESOURCE** | **SOURCE** | **IDENTIFIER** |
| --- | --- | --- |
| **Deposited Data** | | |
| Benchmark scRNA-seq data (Biase) | [Biase et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib4) | GEO: [GSE57249](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57249) |
| Benchmark scRNA-seq data (Treutlein) | [Treutlein et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib39) | Supplementary Data 3 |
| Benchmark scRNA-seq data (Goolam) | [Goolam et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib10) | ArrayExpress: E-MTAB-3321 |
| Benchmark scRNA-seq data (Ting) | [Ting et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib38) | GEO: [GSE51372](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51372) |
| Benchmark scRNA-seq data (Buettner) | [Buettner et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib6) | ArrayExpress: E-MTAB-2805 |
| Benchmark scRNA-seq data (Deng) | [Deng et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib7) | GEO: [GSE45719](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45719) |
| Benchmark scRNA-seq data (Pollen) | [Pollen et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib29) | SRA: SRP041736 |
| Benchmark scRNA-seq data (Patel) | [Patel et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib26) | GEO: [GSE57872](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57872) |
| Benchmark scRNA-seq data (Usoskin) | [Usoskin et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib41) | GEO: [GSE59739](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59739) |
| Benchmark scRNA-seq data (Kolodziejczyk) | [Kolodziejczyk et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib20) | ArrayExpress: E-MTAB-2600 |
| Benchmark scRNA-seq data (Klein) | [Klein et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib19) | GEO: [GSE65525](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525) |
| Benchmark scRNA-seq data (Zeisel) | [Zeisel et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib48) | GEO: [GSE60361](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361) |
| Benchmark scRNA-seq data (PBMC68k) | [Zheng et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib49) | <https://support.10xgenomics.com/single-cell-gene-expression/datasets> |
| Benchmark scRNA-seq data (Brain1m) | 10x Genomics | <https://support.10xgenomics.com/single-cell-gene-expression/datasets> |
| Benchmark scRNA-seq data (FACS-purified blood cells) | [Zheng et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib49) | <https://support.10xgenomics.com/single-cell-gene-expression/datasets> |
| **Software and Algorithms** | | |
| Barnes-Hut t-SNE | [Van Der Maaten, 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib42) | <https://github.com/lvdmaaten/bhtsne> |
| Scikit-learn (k-means and agglomerative clustering) | [Pedregosa et al., 2011](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib27) | <http://scikit-learn.org/stable/>; RRID:SCR\_002577 |
| **Other** | | |
| Software package for net-SNE | This paper | [http://netsne.csail.mit.edu](http://netsne.csail.mit.edu/) |

Contact for Reagent and Resource Sharing

Bonnie Berger, [bab@mit.edu](mailto:bab@mit.edu), Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.

Method Details

Review of t-Stochastic Neighbor Embedding

Let x1,…,xn∈Rd represent the (normalized) [expression profiles](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gene-expression-profiling) for each of the *n* cells in a scRNA-seq dataset that we wish to visualize in Rs, where *d* is typically on the order of tens of thousands (number of human genes) and *s* is two or three. More precisely, we want to learn the low-dimensional embedding of the cells y1,…,yn∈Rs that capture the low-dimensional structure represented by the original input vectors ***x***1,…,***x****n*.

A widely-used approach called t-stochastic neighbor embedding (t-SNE) ([Maaten and Hinton, 2008](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib23)) relates the notion of *quality* of an embedding ***y***1,…,***y****n* (inversely) to the Kullback-Leibler (KL) divergence between the two [probability distributions](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/probability-distribution) ***P*** and ***Q*** defined over all pairs of cells, which reflect how the cells are laid out in the input and output (embedding) spaces, respectively. The probability assigned to a particular pair (*i*, *j*) in each distribution represents how close the two associated vectors are---i.e., (***x****i*, ***x****j*) for ***P*** and (***y****i*, ***y****j*) for ***Q***. Intuitively, maximizing the agreement between ***P*** and ***Q*** corresponds to finding a good embedding that faithfully represents the structure in the original data.

Formally, t-SNE solves the following optimization problem

minimizey1,…,ynKL(P∥Q)=∑i≠jpijlogpijqij,

where

pij=pi|j+pi|j2withpi|j=exp{−‖xi−xj‖22/(2σj2)}∑i′≠jexp{−‖xi′−xj‖22/(2σj2)},

and

qij=q˜ijZwithq˜ij=(1+‖yi−yj‖22)−1andZ=∑i≠jq˜ij.

Note *pij* and *qij* denote the (*i*, *j*) element of matrices ***P*** and ***Q***, respectively. In addition, *σj* is a parameter that is tuned for each *j* to ensure pi|j achieves a predefined value of information-theoretic entropy.

t-SNE solves the above optimization problem via gradient descent on the embedding vectors ***y***1,…,***y****n* with random initialization. As derived in the original paper ([Maaten and Hinton, 2008](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib23)), the gradient of the objective with respect to each ***y****i* is given as

δKL(P∥Q)δyi=∑j≠ipijq˜ij(yi−yj)−1Z∑j≠iq˜ij2(yi−yj).

Computing this gradient for every cell *i* would require *O*(*n*2) computation, which is prohibitive for large *n*. In the state-of-the-art implementation of t-SNE ([Van Der Maaten, 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib42)), this expression is approximated in two ways for computational efficiency. First, ***P*** is approximated with a sparse matrix based on k-nearest neighbors for each cell, which greatly speeds up the computation of the first term since most summands are zero. Second, an efficient data structure (space-partitioning trees ([Samet, 1984](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib33))) is built over ***y***1,…,***y****n* so that the second summation can be coarsely approximated by grouping terms corresponding to nearby ***y****i*'s together. Even with these optimizations, applying t-SNE to datasets with millions of cells requires days of computation as shown in our results.

Our Method: Neural t-SNE

We introduce neural t-SNE (net-SNE), which models each embedding vector ***y****i* as the output of a parameterized embedding function evaluated at the corresponding input vector ***x****i*. Importantly, our approach is generalizable—i.e., it induces the embedding of any point in the input space, not just the observed data points as in t-SNE. We use standard feedforward neural networks (NNs) ([LeCun et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib21)) to represent the embedding function, drawing from the intuition that NNs have sufficient expressive capacity to find high-quality maps similar to those typically uncovered by t-SNE.

The precise form of the parameterized mapping of net-SNE is as follows. Let *ℓ* be the number of hidden layers in the NN and *u* be the number of units in each layer (same for every layer). Furthermore, let W(t)∈Ru×u (Ru×d for *t* = 1) be the weight matrix and b(t)∈Ru be the intercept associated with layers *t* = 1,…,*ℓ*. An additional weight matrix ***W***(*ℓ*+1)is associated with the final output layer. Given a data point ***x****i*, the forward pass through the NN to compute the embedding ***y****i* can be recursively described as

hi(0)=xi,hi(t)=f(W(t)hi(t−1)+b(t)),fort=1,…,ℓ,yi=W(ℓ+1)hi(ℓ).

Note that *f* denotes an element-wise nonlinear activation function (e.g., sigmoid or rectifier). In the following, we compactly represent the above NN-based embedding function as yi=NN(xi;Θ), where Θ refers to the network parameters ***W***(1),…,***W***(*ℓ*+1) and ***b***(1),…,***b***(*ℓ*). All of our experimental results are obtained uniformly based on a simple architecture with *ℓ* = 2, *u* = 50, and rectifier activation, as illustrated in [Figure S1](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)B.

Given a NN that defines an embedding for every point in the input space, net-SNE optimizes the same KL divergence objective as t-SNE over the observed data points, via gradient descent. To see how the gradients are computed in net-SNE, first note that for a particular network parameter θ∈Θ, we have

δKL(P∥Q)δθ=∑i=1n(δKL(P∥Q)δyi)TδNN(xi;Θ)δθ

by the chain rule and using yi=NN(xi;Θ). Notice the first term in each product is identical to the t-SNE gradient and can be computed in the same manner. The second term can be computed via standard backpropagation algorithm ([LeCun et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib21)).

Intuitively, here we keep most of the computation in t-SNE intact, but add an additional step to each iteration, where, after computing the gradients for ***y***1,…,***y****n* as in t-SNE, we propagate them backward through the NN to update the network weights accordingly. Consequently, net-SNE is compatible with any computational optimization for t-SNE. In particular, our implementation of net-SNE incorporates the state-of-the-art version of t-SNE based on the Barnes-Hut approximation ([Van Der Maaten, 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib42)) to the gradients with respect to ***y***1,…,***y****n*, which achieves substantially faster runtime than vanilla t-SNE.

Although our method was independently developed, we note that a theoretical approach of training a parametric embedding (e.g., a neural network) for t-SNE via the chain rule has been previously described in an earlier work ([Van Der Maaten, 2009](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib43)). However, given the difficulty in successfully training a neural network to find good solutions to the t-SNE objective on large-scale datasets, practical adoption of this approach has been limited. In our work, we introduce additional techniques described in the following sections to improve the effectiveness of neural network-based visualization, while also demonstrating its utility for [single cell analysis](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/single-cell-analysis) on a wide range of benchmark datasets.

Accelerating net-SNE via Stochastic Optimization

To fully exploit the generalizability of net-SNE, we improve upon the above procedure with techniques from stochastic optimization. First, note that the gradient for θ∈Θ given in the previous section is a summation over all data points, and thus can be approximated with a randomly chosen subset B⊂{1,…,n} (“mini-batch”) as

δKL(P∥Q)δθ≈n|B|∑i∈B(δKL(P∥Q)δyi)TδNN(xi;Θ)δθ.

Similarly, the gradient with respect to ***y****i* for *i* ∈ B can be approximated as

δKL(P∥Q)δyi≈n−1|B|−1∑j∈B∧j≠ipijq˜ij(yi−yj)−n−1|B−1|⋅1Zˆ∑j∈B∧j≠iq˜ij2(yi−yj)

using the approximate normalization factor

Zˆ=n|B|⋅n−1|B|−1∑i,j∈B∧i≠jqˆij.

With small |B| and precomputed ***P*** (which typically constitutes a small fraction of the runtime of t-SNE), this approach greatly reduces the required computation in each gradient step and improves the rate of convergence to a good visualization, as we demonstrate in our experiments. Our observation is in line with well-known results in the field of optimization ([Bousquet and Bottou, 2008](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "bib5)) showing superior runtimes of stochastic gradient descent (SGD) methods compared to their exact counterparts that process the entire dataset for every iteration. Notably, although only a small subset of cells are considered for each iteration, each of our gradient updates to Θ affects all cells in the data (due to its generalizability). In contrast, applying a similar mini-batch SGD procedure to t-SNE results in an ineffective method, as only the positions of the cells in a given mini-batch are updated while the remaining cells are fixed. This reduces to a coordinate descent-like procedure, which we found to be very slow in terms of learning speed, likely due to the tight coupling of parameters being optimized. Although net-SNE shares the model parameters across all cells and thus is less prone to this issue, we did notice difficulties in optimization with mini-batches that are too small. We found setting |B| to be around 10% of the dataset to be a reasonable compromise that leads to good performance.

In addition, sampling strategy for B has a considerable effect on the quality of approximation for the gradients. Specifically, given a sparse approximation of ***P***, the first summation in our equation for δKL(P∥Q)/δyi has only a few nonzero summands. If B is uniformly sampled, then only a few indices will contribute to the sum, leading to high variance in the estimate. We address this problem by introducing additional structure into B. In particular, we first sample a smaller set of seed cells S⊂{1,…,n} uniformly at random and then sample a fixed number of cells from each of their “neighbors” (where *pij* > 0). After these local samples are used to facilitate the approximation of t-SNE gradients for the seed cells, these gradients are backpropagated through the neural network to update the embedding parameters of net-SNE. Given that our estimated normalization factor Zˆ appears in the denominator of the approximate t-SNE gradient, our gradient estimate based on Zˆ is thus biased. To control the amount of error introduced, we impose a minimum threshold (10%) on the fraction of total samples used to approximate Zˆ. If the set of seed cells is too small, additional samples are drawn to ensure Zˆ is of sufficient quality.

Training net-SNE with Reference Visualization

Even with our stochastic optimization techniques, training a neural network to optimize the t-SNE objective is a challenging task, especially for large-scale datasets with complex patterns. We thus introduce another technique, where a pre-trained t-SNE embedding is used to provide more direct feedback to the neural network instead of relying on the gradients of the highly complex t-SNE objective. More precisely, mean-squared error between the net-SNE embedding and an existing t-SNE map is used as the loss function to optimize the network in order to obtain a good initial solution, which can be further optimized if needed. Although this does require t-SNE to be performed before applying net-SNE, the initial t-SNE map need not be fully optimized, as further SGD iterations in net-SNE can fine-tune the solution.

Note that all of our datasets other than PBMC68k and Brain1m are small enough that we were able to train net-SNE in batch mode without the stochastic optimization and the use of a reference visualization. On the other hand, our results on PBMC68k and the 100k subset of Brain1m were obtained by training net-SNE to match a t-SNE reference using mini-batch SGD with a batch size of 10%. We did not further fine-tune net-SNE after training on the reference visualization, as the resulting visualization quality was sufficiently high. After generalizing the 100k cell visualization to the full Brain1m dataset, further training of net-SNE was performed without a t-SNE map (since t-SNE becomes impractical at this scale); instead, we used our stochastic optimization with a batch size of 10%. Although our generalization based on a smaller subset is the primary factor in achieving fast runtime for the Brain1m visualization, our stochastic optimization techniques also lead to significant runtime reductions. For instance, with a batch size of 10%, the average runtime of each iteration of net-SNE is reduced by around a factor of 10 as a result of our techniques.

Benchmark Datasets

For our main experiments, we used 13 published scRNA-seq datasets of varying sizes with known cluster labels for the cells, which allowed us to directly assess the quality of the visualization produced by net-SNE. The list of datasets, sorted in increasing order of size: ([Biase et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib4)) (*n* = 49, *k* = 3), ([Treutlein et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib39)) (*n* = 80, *k* = 5), ([Goolam et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib10)) (*n* = 124, *k* = 5), ([Ting et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib38)) (*n* = 149, *k* = 7), ([Buettner et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib6)) (*n* = 182, *k* = 3), ([Deng et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib7)) (*n* = 268, *k* = 10), ([Pollen et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib29)) (*n* = 301, *k* = 11), ([Patel et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib26)) (*n* = 430, *k* = 5), ([Usoskin et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib41)) (*n* = 622, *k* = 4), ([Kolodziejczyk et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib20)) (*n* = 704, *k* = 3), ([Klein et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib19)) (*n* = 2,717, *k* = 4), ([Zeisel et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib48)) (*n* = 3,005, *k* = 9), and PBMC68k ([Zheng et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib49)) (*n* = 68,560, *k* = 10). Note that *n* denotes the number of cells and *k* denotes the number of clusters defined by the original publications.

We additionally used a mega-scale dataset Brain1m (*n* = 1,283,543), downloaded from the 10x Genomics website (<https://support.10xgenomics.com/single-cell-gene-expression/datasets>), to assess the generalizability and scalability of net-SNE. However, this dataset does not have any known labels, and thus we resorted to using the objective score optimized by both net-SNE and t-SNE as the quality metric for comparison.

For demonstrating the generalization of net-SNE across different datasets, we used the datasets of six blood cell subtypes (CD4+, [CD8+](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/cytotoxic-t-cell), [CD14+](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/cd14), CD19+, CD34+, and CD56+) experimentally purified via [fluorescence activated cell sorting](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/flow-cytometry) (FACS), which are provided by 10x Genomics and described in [Zheng et al. (2017)](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bib49). These datasets were downloaded from the same URL as the Brain1m dataset given above.

Our benchmark datasets measured gene expression via a number of different metrics, including read, fragment, transcript, or unique molecule (UMI) counts that are either normalized or unnormalized for gene length ([Table S1](https://www.sciencedirect.com/science/article/pii/S2405471218302357#mmc1)). We kept the metric chosen by the original publication for each dataset with the goal of demonstrating the performance of net-SNE in a variety of settings. Unless already performed by the original publication, we applied the standard *log*(1 + *x*) transformation to each element *x* of the cell-gene expression matrix before visualizing the data.

Data and Software Availability

A C++ implementation of net-SNE with example data and scripts are available at: [http://netsne.csail.mit.edu](http://netsne.csail.mit.edu/) and <https://github.com/hhcho/netsne>. The accession numbers for all scRNA-seq datasets analyzed in this paper are provided in [Key Resources Table](https://www.sciencedirect.com/science/article/pii/S2405471218302357" \l "sec4.1). Preprocessed gene-expression matrices are available from the authors upon request.

Acknowledgments

H.C. and B.B. are partially supported by the US NIH grant [R01GM081871](https://www.sciencedirect.com/science/article/pii/S2405471218302357#gs1) (to B.B.). H.C. is also partially supported by the Kwanjeong Educational Foundation. J.P. is supported by the Sloan Research Fellowship and the US National Science Foundation Career Award 1652815. We thank Serafim Batzoglou and Bo Wang for providing the preprocessed data for Pollen and Kolodziejczyk datasets. Editor's note: An early version of this paper was submitted to and peer reviewed at the 2018 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at *Cell Systems*.

Author Contributions

Conceptualization, H.C., B.B., and J.P.; Data Curation, H.C. and J.P.; Investigation, H.C., B.B., and J.P.; Methodology, H.C., B.B., and J.P.; Resources, B.B.; Software, H.C.; Validation, H.C.; Visualization, H.C.; Writing – Original Draft, H.C.; Writing – Review & Editing, H.C., B.B., and J.P.; Funding Acquisition, B.B.; Supervision, B.B. and J.P.

Declaration of Interests

The authors declare no conflicting interests.

Supplemental Information

[Download Acrobat PDF file (1MB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-mmc1.pdf)[Help with pdf files](https://service.elsevier.com/app/answers/detail/a_id/19286/supporthub/sciencedirect/)

Document S1. Figures S1–S4 and Table S1.

[Download Acrobat PDF file (4MB)](https://ars.els-cdn.com/content/image/1-s2.0-S2405471218302357-mmc2.pdf)[Help with pdf files](https://service.elsevier.com/app/answers/detail/a_id/19286/supporthub/sciencedirect/)

Document S2. Article plus Supplemental Information.

References

[Amir et al., 2013](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib1)

el-A.D. Amir, K.L. Davis, M.D. Tadmor, E.F. Simonds, J.H. Levine, S.C. Bendall, D.K. Shenfeld, S. Krishnaswamy, G.P. Nolan, D. Pe'er**viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia**

Nat. Biotechnol., 31 (2013), p. 545

[CrossRef](https://doi.org/10.1038/nbt.2594)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84880280631&partnerID=10&rel=R3.0.0)

[Amodio et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib2)

M. Amodio, K. Srinivasan, D. van Dijk, H. Moshen, K. Yim, R. Muhle, K.R. Moon, S. Kaech, R. Sowell, R. Montgomery, *et al.***Exploring single-cell data with deep multitasking neural networks**

bioRxiv (2017), [10.1101/237065](https://doi.org/10.1101/237065)

[Anchang et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib3)

B. Anchang, T.D.P. Hart, S.C. Bendall, P. Qiu, Z. Bjornson, M. Linderman, G.P. Nolan, S.K.Plevritis**Visualization and cellular hierarchy inference of single-cell data using SPADE**

Nat. Protoc., 11 (2016), pp. 1264-1279

[CrossRef](https://doi.org/10.1038/nprot.2016.066)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84975822112&partnerID=10&rel=R3.0.0)

[Biase et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib4)

F.H. Biase, X. Cao, S. Zhong**Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing**

Genome Res., 24 (2014), pp. 1787-1796

[CrossRef](https://doi.org/10.1101/gr.177725.114)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84928227321&partnerID=10&rel=R3.0.0)

[Bousquet and Bottou, 2008](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib5)

O. Bousquet, L. Bottou**The tradeoffs of large scale learning**

D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), Advances in Neural Information Processing Systems 21, NIPS (2008), pp. 161-168

[CrossRef](https://doi.org/10.1109/ICST.2008.34)

[Buettner et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib6)

F. Buettner, K.N. Natarajan, F.P. Casale, V. Proserpio, A. Scialdone, F.J. Theis, S.A.Teichmann, J.C. Marioni, O. Stegle**Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells**

Nat. Biotechnol., 33 (2015), pp. 155-160

[CrossRef](https://doi.org/10.1038/nbt.3102)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84923292191&partnerID=10&rel=R3.0.0)

[Deng et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib7)

Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg**Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells**

Science, 343 (2014), pp. 193-196

[CrossRef](https://doi.org/10.1126/science.1245316)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84892179132&partnerID=10&rel=R3.0.0)

[Dzwinel and Wcisło, 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib8)

W. Dzwinel, R. Wcisło**Very fast interactive visualization of large sets of high-dimensional data**

Procedia Comput. Sci., 51 (2015), pp. 572-581

[Article](https://www.sciencedirect.com/science/article/pii/S1877050915011333)[Download PDF](https://www.sciencedirect.com/science/article/pii/S1877050915011333/pdf?md5=b792c6fcb9940289b0f498c471353189&pid=1-s2.0-S1877050915011333-main.pdf)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84939178884&partnerID=10&rel=R3.0.0)

[Gawad et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib9)

C. Gawad, W. Koh, S.R. Quake**Single-cell genome sequencing: current state of the science**

Nat. Rev. Genet., 17 (2016), pp. 175-188

[CrossRef](https://doi.org/10.1038/nrg.2015.16)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84958606331&partnerID=10&rel=R3.0.0)

[Goolam et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib10)

M. Goolam, A. Scialdone, S.J. Graham, I.C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J.C. Marioni, M. Zernicka-Goetz**Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos**

Cell, 165 (2016), pp. 61-74

[Article](https://www.sciencedirect.com/science/article/pii/S0092867416300617)[Download PDF](https://www.sciencedirect.com/science/article/pii/S0092867416300617/pdfft?md5=1551d66d138072bac9e5e8683d68349e&pid=1-s2.0-S0092867416300617-main.pdf)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84961775163&partnerID=10&rel=R3.0.0)

[Grün et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib11)

D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. van Oudenaarden**Single-cell messenger RNA sequencing reveals rare intestinal cell types**

Nature, 525 (2015), pp. 251-255

[CrossRef](https://doi.org/10.1038/nature14966)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84941201582&partnerID=10&rel=R3.0.0)

[Haghverdi et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib12)

L. Haghverdi, A.T.L. Lun, M.D. Morgan, J.C. Marioni**Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours**

bioRxiv (2017), [10.1101/165118](https://doi.org/10.1101/165118)

[Hartigan and Wong, 1979](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib13)

J.A. Hartigan, M.A. Wong**Algorithm AS 136: a k-means clustering algorithm**

J. R. Stat. Soc. Ser. C Appl. Stat., 28 (1979), pp. 100-108

[CrossRef](https://doi.org/10.2307/2346830)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84873576658&partnerID=10&rel=R3.0.0)

[Hutchison et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib14)

L.A.D. Hutchison, B. Berger, I. Kohane***C. elegans* exhibits coordinated oscillation in gene expression during development**

bioRxiv (2017), [10.1101/114074](https://doi.org/10.1101/114074)

[Jackson, 2005](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib15)

J.E. Jackson**A User's Guide to Principal Components**

John Wiley & Sons (2005)

[Jaitin et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib16)

D.A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N.Cohen, S. Jung, A. Tanay, *et al.***Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types**

Science, 343 (2014), pp. 776-779

[CrossRef](https://doi.org/10.1126/science.1247651)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84893905629&partnerID=10&rel=R3.0.0)

[Kikuchi-Taura et al., 2006](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib17)

A. Kikuchi-Taura, T. Soma, T. Matsuyama, D.M. Stern, A. Taguchi**A new protocol for quantifying CD34+ cells in peripheral blood of patients with cardiovascular disease**

Tex. Heart Inst. J., 33 (2006), p. 427

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-33846227512&partnerID=10&rel=R3.0.0)

[Kiselev et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib18)

V.Y. Kiselev, K. Kirschner, M.T. Schaub, T. Andrews, A. Yiu, T. Chandra, K.N. Natarajan, W.Reik, M. Barahona, A.R. Green, *et al.***SC3: consensus clustering of single-cell RNA-seq data**

Nat. Methods, 14 (2017), pp. 483-486

[CrossRef](https://doi.org/10.1038/nmeth.4236)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85016121177&partnerID=10&rel=R3.0.0)

[Klein et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib19)

A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner**Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells**

Cell, 161 (2015), pp. 1187-1201

[Article](https://www.sciencedirect.com/science/article/pii/S0092867415005000)[Download PDF](https://www.sciencedirect.com/science/article/pii/S0092867415005000/pdfft?md5=d790935fd967fbfab97b6a7104a628eb&pid=1-s2.0-S0092867415005000-main.pdf)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84929684998&partnerID=10&rel=R3.0.0)

[Kolodziejczyk et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib20)

A.A. Kolodziejczyk, J.K. Kim, J.C. Tsang, T. Ilicic, J. Henriksson, K.N. Natarajan, A.C.Tuck, X. Gao, M. Bühler, P. Liu, *et al.***Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation**

Cell Stem Cell, 17 (2015), pp. 471-485

[Article](https://www.sciencedirect.com/science/article/pii/S193459091500418X)[Download PDF](https://www.sciencedirect.com/science/article/pii/S193459091500418X/pdfft?md5=fc07d0bc92113f84bbc7717fd59fa424&pid=1-s2.0-S193459091500418X-main.pdf)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84947748539&partnerID=10&rel=R3.0.0)

[LeCun et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib21)

Y. LeCun, Y. Bengio, G. Hinton**Deep learning**

Nature, 521 (2015), pp. 436-444

[CrossRef](https://doi.org/10.1038/nature14539)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84930630277&partnerID=10&rel=R3.0.0)

[Loh et al., 2012](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib22)

P.-R. Loh, M. Baym, B. Berger**Compressive genomics**

Nat. Biotechnol., 30 (2012), pp. 627-630

[CrossRef](https://doi.org/10.1038/nbt.2241)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84863702145&partnerID=10&rel=R3.0.0)

[Maaten and Hinton, 2008](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib23)

L.V.D. Maaten, G. Hinton**Visualizing data using t-SNE**

J. Mach. Learn. Res., 9 (2008), pp. 2579-2605

[Moon et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib24)

K.R. Moon, D. van Dijk, Z. Wang, D. Burkhardt, W. Chen, A. van den Elzen, M.J. Hirn, R.R.Coifman, N.B. Ivanova, G. Wolf, S. Krishnaswamy**Visualizing transitions and structure for high dimensional data exploration**

bioRxiv (2017), [10.1101/120378](https://doi.org/10.1101/120378)

[Palmer et al., 2012](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib25)

N.P. Palmer, P.R. Schmid, B. Berger, I.S. Kohane**A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers**

Genome Biol., 13 (2012), p. R71

[CrossRef](https://doi.org/10.1186/gb-2012-13-8-r71)

[Patel et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib26)

A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V.Nahed, W.T. Curry, R.L. Martuza, *et al.***Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma**

Science, 344 (2014), pp. 1396-1401

[CrossRef](https://doi.org/10.1126/science.1254257)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84902668801&partnerID=10&rel=R3.0.0)

[Pedregosa et al., 2011](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib27)

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.Prettenhofer, R. Weiss, V. Dubourg, *et al.***Scikit-learn: machine learning in Python**

J. Mach. Learn. Res., 12 (2011), pp. 2825-2830

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-80555140075&partnerID=10&rel=R3.0.0)

[Pierson and Yau, 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib28)

E. Pierson, C. Yau**ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis**

Genome Biol., 16 (2015), p. 241

[Pollen et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib29)

A.A. Pollen, T.J. Nowakowski, J. Shuga, X. Wang, A.A. Leyrat, J.H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, *et al.***Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex**

Nat. Biotechnol., 32 (2014), pp. 1053-1058

[CrossRef](https://doi.org/10.1038/nbt.2967)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84922321862&partnerID=10&rel=R3.0.0)

[Qiu et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib30)

X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, C. Trapnell**Reversed graph embedding resolves complex single-cell trajectories**

Nat. Methods, 14 (2017), pp. 979-982

[CrossRef](https://doi.org/10.1038/nmeth.4402)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85031017685&partnerID=10&rel=R3.0.0)

[Rand, 1971](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib31)

W.M. Rand**Objective criteria for the evaluation of clustering methods**

J. Am. Stat. Assoc., 66 (1971), pp. 846-850

[CrossRef](https://doi.org/10.1080/01621459.1971.10482356)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84950632109&partnerID=10&rel=R3.0.0)

[Regev et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib32)

A. Regev, S. Teichmann, E.S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P.Campbell, P. Carninci, M. Clatworthy, *et al.***The human cell atlas**

Elife, 6 (2017), [10.7554/eLife.27041](https://doi.org/10.7554/eLife.27041)

[Samet, 1984](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib33)

H. Samet**The quadtree and related hierarchical data structures**

ACM Comput. Surv., 16 (1984), pp. 187-260

[CrossRef](https://doi.org/10.1145/356924.356930)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-0021644214&partnerID=10&rel=R3.0.0)

[Simmons et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib34)

S. Simmons, J. Peng, J. Bienkowska, B. Berger**Discovering what dimensionality reduction really tells us about RNA-seq data**

J. Comput. Biol., 22 (2015), pp. 715-728

[CrossRef](https://doi.org/10.1089/cmb.2015.0085)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84938383708&partnerID=10&rel=R3.0.0)

[Stubbington et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib35)

M.J. Stubbington, O. Rozenblatt-Rosen, A. Regev, S.A. Teichmann**Single-cell transcriptomics to explore the immune system in health and disease**

Science, 358 (2017), pp. 58-63

[CrossRef](https://doi.org/10.1126/science.aan6828)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85030648051&partnerID=10&rel=R3.0.0)

[Svensson et al., 2018](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib36)

V. Svensson, R. Vento-Tormo, S.A. Teichmann**Exponential scaling of single-cell RNA-seq in the past decade**

Nat. Protoc., 13 (2018), pp. 599-604

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85044252958&partnerID=10&rel=R3.0.0)

[Tang et al., 2016](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib37)

Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016). Visualizing large-scale and high-dimensional data. Proceedings of the 25th International Conference on World Wide Web 287–297. <https://doi.org/10.1145/2872427.2883041>.

[Ting et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib38)

D.T. Ting, B.S. Wittner, M. Ligorio, N.V. Jordan, A.M. Shah, D.T. Miyamoto, N. Aceto, F. Bersani, B.W. Brannigan, K. Xega, *et al.***Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells**

Cell Rep., 8 (2014), pp. 1905-1918

[Article](https://www.sciencedirect.com/science/article/pii/S2211124714007050)[Download PDF](https://www.sciencedirect.com/science/article/pii/S2211124714007050/pdfft?md5=022db5f6dc3adb25ed219839913bdc49&pid=1-s2.0-S2211124714007050-main.pdf)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84907414338&partnerID=10&rel=R3.0.0)

[Treutlein et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib39)

B. Treutlein, D.G. Brownfield, A.R. Wu, N.F. Neff, G.L. Mantalas, F.H. Espinoza, T.J. Desai, M.A. Krasnow, S.R. Quake**Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq**

Nature, 509 (2014), pp. 371-375

[CrossRef](https://doi.org/10.1038/nature13173)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84900529199&partnerID=10&rel=R3.0.0)

[Tung et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib40)

P.-Y. Tung, J.D. Blischak, C.J. Hsiao, D.A. Knowles, J.E. Burnett, J.K. Pritchard, Y. Gilad**Batch effects and the effective design of single-cell gene expression studies**

Sci. Rep., 7 (2017), p. 39921

[CrossRef](https://doi.org/10.1038/srep39921)

[Usoskin et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib41)

D. Usoskin, A. Furlan, S. Islam, H. Abdo, P. Lönnerberg, D. Lou, J. Hjerling-Leffler, J.Haeggström, O. Kharchenko, P.V. Kharchenko, *et al.***Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing**

Nat. Neurosci., 18 (2015), pp. 145-153

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84921466417&partnerID=10&rel=R3.0.0)

[Van Der Maaten, 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib42)

L. Van Der Maaten**Accelerating t-SNE using tree-based algorithms**

J. Mach. Learn. Res., 15 (2014), pp. 3221-3245

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84919775831&partnerID=10&rel=R3.0.0)

[Van Der Maaten, 2009](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib43)

L. Van Der Maaten**Learning a parametric embedding by preserving local structure**

RBM, 500 (2009), p. 26

[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85020752494&partnerID=10&rel=R3.0.0)

[Wang et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib44)

B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou**Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning**

Nat. Methods, 14 (2017), pp. 414-416

[CrossRef](https://doi.org/10.1038/nmeth.4207)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85014528252&partnerID=10&rel=R3.0.0)

[Wang et al., 2014](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib45)

Y. Wang, J. Waters, M.L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H.Liang, *et al.***Clonal evolution in breast cancer revealed by single nucleus genome sequencing**

Nature, 512 (2014), pp. 155-160

[Article](https://www.sciencedirect.com/science/article/pii/S0929664612002719)[Download PDF](https://www.sciencedirect.com/science/article/pii/S0929664612002719/pdfft?md5=df9840e09b7ef1eb3e66e6b7b1251de7&pid=1-s2.0-S0929664612002719-main.pdf)[CrossRef](https://doi.org/10.1038/nature13600)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84895873587&partnerID=10&rel=R3.0.0)

[Yoon et al., 2011](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib46)

H.S. Yoon, D.C. Price, R. Stepanauskas, V.D. Rajah, M.E. Sieracki, W.H. Wilson, E.C. Yang, S.Duffy, D. Bhattacharya**Single-cell genomics reveals organismal interactions in uncultivated marine protists**

Science, 332 (2011), pp. 714-717

[CrossRef](https://doi.org/10.1126/science.1203163)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-79955787125&partnerID=10&rel=R3.0.0)

[Yu et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib47)

Y.W. Yu, N.M. Daniels, D.C. Danko, B. Berger**Entropy-scaling search of massive biological data**

Cell Syst., 1 (2015), pp. 130-140

[Article](https://www.sciencedirect.com/science/article/pii/S2405471215000587)[Download PDF](https://www.sciencedirect.com/science/article/pii/S2405471215000587/pdfft?md5=a714cb17bda1c4645800353735e9f6cf&pid=1-s2.0-S2405471215000587-main.pdf)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84941648279&partnerID=10&rel=R3.0.0)

[Zeisel et al., 2015](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib48)

A. Zeisel, A.B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S.Marques, H. Munguba, L. He, C. Betsholtz, *et al.***Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq**

Science, 347 (2015), pp. 1138-1142

[CrossRef](https://doi.org/10.1126/science.aaa1934)[View Record in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84924565530&partnerID=10&rel=R3.0.0)

[Zheng et al., 2017](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bbib49)

G.X. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D.Wheeler, G.P. McDermott, J. Zhu, *et al.***Massively parallel digital transcriptional profiling of single cells**

Nat. Commun., 8 (2017), p. 14049

[CrossRef](https://doi.org/10.1038/ncomms14049)

[4](https://www.sciencedirect.com/science/article/pii/S2405471218302357#bfn1)

Lead Contact

© 2018 Published by Elsevier Inc.

Recommended articles

[Network Visualization and Analysis of Spatially Aware Gene Expression Data with InsituNet](https://www.sciencedirect.com/science/article/pii/S2405471218301078)

Cell Systems, Volume 6, Issue 5, 2018, pp. 626-630.e3

[Download PDF](https://www.sciencedirect.com/science/article/pii/S2405471218301078/pdfft?md5=f7b3d1e2366cbcfd1b237288ba50b557&pid=1-s2.0-S2405471218301078-main.pdf)View details

[Breaking New Ground in the Landscape of Single-Cell Analysis](https://www.sciencedirect.com/science/article/pii/S240547121730594X)

Cell Systems, Volume 6, Issue 1, 2018, pp. 5-7

[Download PDF](https://www.sciencedirect.com/science/article/pii/S240547121730594X/pdfft?md5=7894ef8b6901d07e7c6f48e166de089b&pid=1-s2.0-S240547121730594X-main.pdf)View details

[Research Techniques Made Simple: Single-Cell RNA Sequencing and its Applications in Dermatology](https://www.sciencedirect.com/science/article/pii/S0022202X1830109X)

Journal of Investigative Dermatology, Volume 138, Issue 5, 2018, pp. 1004-1009

[Download PDF](https://www.sciencedirect.com/science/article/pii/S0022202X1830109X/pdfft?md5=a31d4ffb5317688be47c313af711395e&pid=1-s2.0-S0022202X1830109X-main.pdf)View details

**1**2Next

Citing articles (0)

Article Metrics

Social Media

* Tweets:7

https://d39af2mgp1pqhg.cloudfront.net/01b30309c878a019f260d2d9b9773b0c/plumx-logo.png

[View details](https://plu.mx/plum/a/?doi=10.1016/j.cels.2018.05.017&theme=plum-sciencedirect-theme&hideUsage=true)

[About ScienceDirect](https://www.elsevier.com/solutions/sciencedirect)[Remote access](https://www.sciencedirect.com/science/activateaccess)[Shopping cart](https://www.sciencedirect.com/science?_ob=ShoppingCartURL&_method=display&md5=3ff44acb300f01481824c54a2973d019)[Contact and support](https://service.elsevier.com/app/contact/supporthub/sciencedirect/)[Terms and conditions](https://www.elsevier.com/legal/elsevier-website-terms-and-conditions)[Privacy policy](https://www.elsevier.com/legal/privacy-policy)

Cookies are used by this site. For more information, visit the [cookies page](https://www.elsevier.com/solutions/sciencedirect/support/cookies).

Copyright © 2018 Elsevier B.V. or its licensors or contributors. ScienceDirect ® is a registered trademark of Elsevier B.V.