

# t-SNE

## What is t-SNE?

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensional reduction technique [1]. t-SNE aims to preserve the essential high-dimensional structure and present it in a low-dimensional representation. t-SNE is particularly useful for visually identifying groups of similar samples or cells in large high-dimensional data sets such as single cell RNA-Seq.

## Running t-SNE

We recommend normalizing your data prior to running t-SNE, but the task will run on any counts data node.

- Click the counts data node
- Click the **Exploratory analysis** section of the toolbox
- Click **t-SNE**
- Click **Finish** to run

t-SNE produces a t-SNE task node. Opening the task report launches a scatter plot showing the t-SNE results. Each point on the plot is a cell for single cell data or a sample for bulk data. The plot will open in 2D or 3D depending on the user preference.

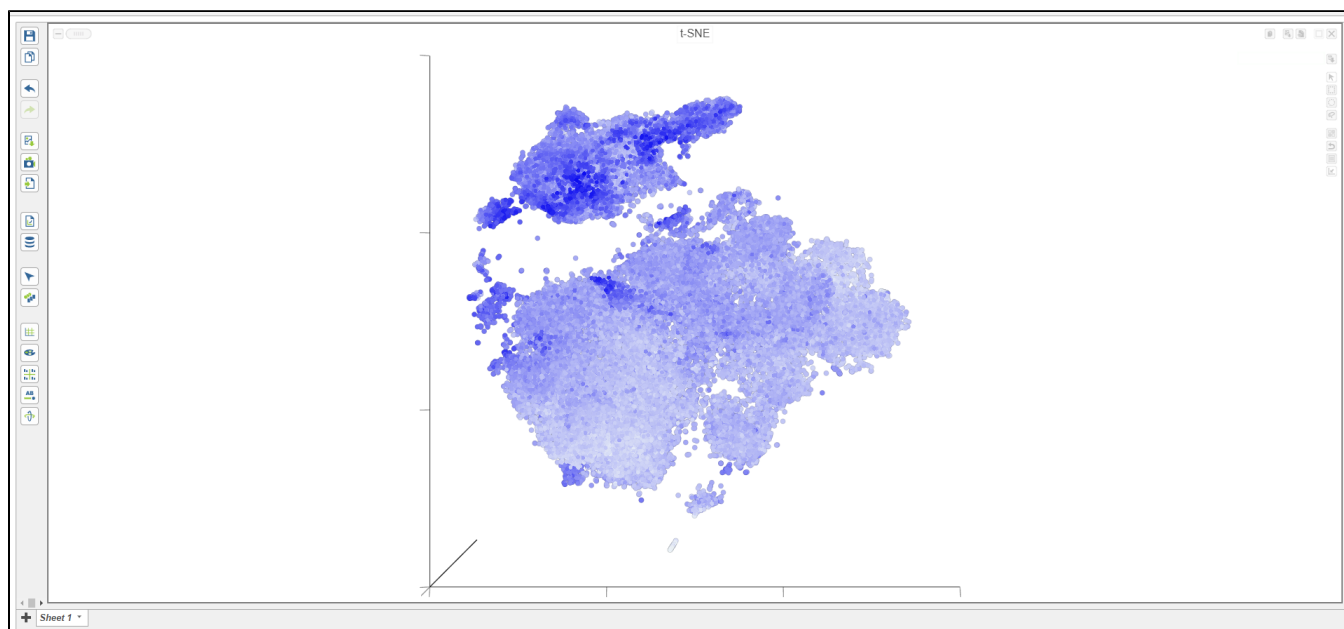


Figure 1. t-SNE task report with interactive scatter plot

## Basic t-SNE parameters

### Split cells by sample

Chose whether to run t-SNE on all samples together or on each sample individually.

Checking the box will run t-SNE on each sample individually.

### Include features where "Feature type" is

This option appears when there are multiple feature types in the input data node (e.g., CITE-Seq data).

Select *Any* to run on all features or pick a feature type.

## Advanced t-SNE parameters

### Perplexity

t-SNE preserves the local structure of the data by focusing on the distances between each point and its nearest neighbors. Perplexity can be thought of as the number of nearest neighbors being considered. The optimal perplexity depends on the size and density of the data. Generally, a larger and/or more dense data set will benefit from a higher perplexity (Figure 2). Default is 30. The range of possible values is 3 to 100.

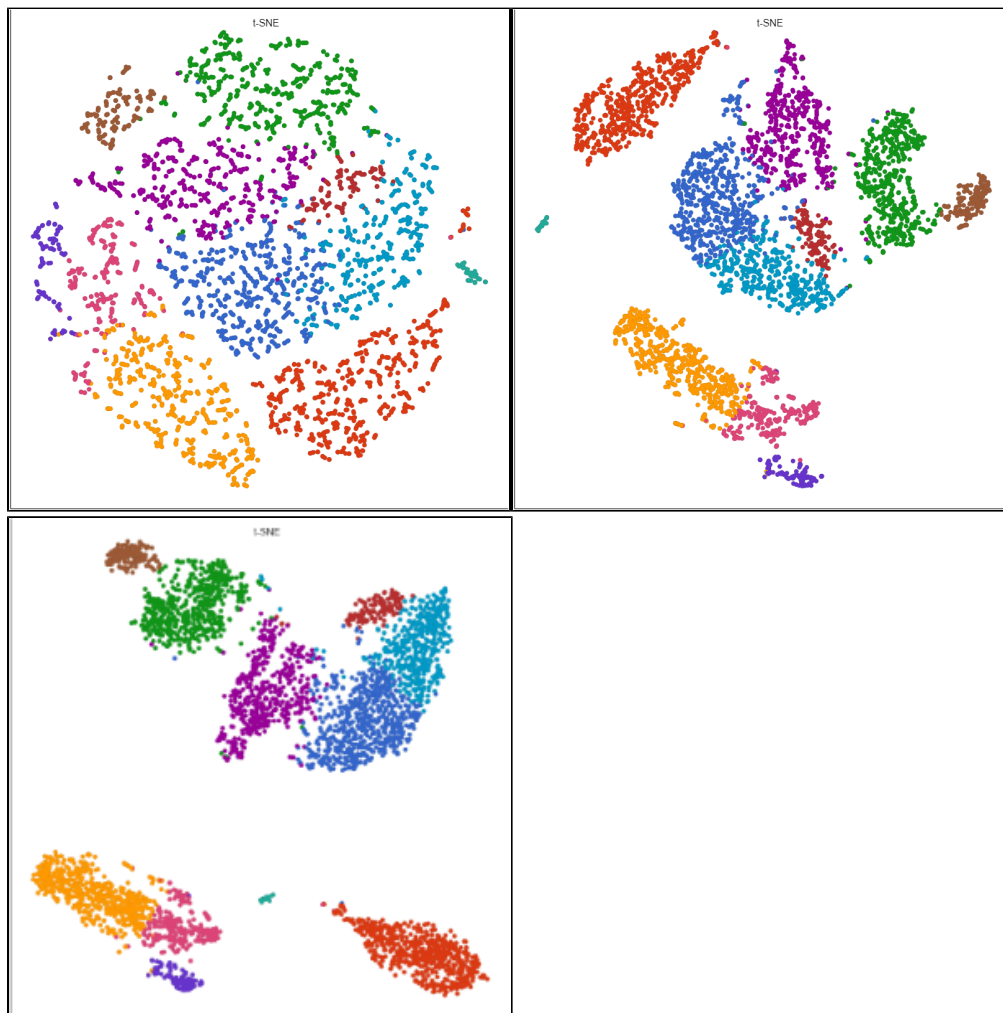


Figure 2. Setting t-SNE perplexity to 5 (left), 30 (middle), 100 (right)

### Number of iterations

t-SNE uses an iterative algorithm to optimize the low-dimensional representation. More iterations will result in a more accurate embedding to an extent, but will take longer to run. Default is 1000.

### Random generator seed

Several parts of t-SNE utilize a random number generator to provide an initial value. Default is 1. To reproduce the results, use the same random seed at all runs.

### Initialize output values at random

If selected, t-SNE initializes from random initial positions for each point. If disabled, the initial values for each point are assigned using the largest principal components extracted from the raw data. Default is enabled.

### Distance metric

The metric to use when computing distances in high-dimensional space. Options are *Euclidean*, *Manhattan*, *Chebyshev*, *Canberra*, *Bray Curtis*, and *Cosine*. Default is *Euclidean*.

### Generate mapping error statistics

If checked, mapping error information will be available in the task report. Default is disabled.

## Generate t-SNE table

Output a t-SNE table data node that can be downloaded. The 2D t-SNE coordinates are labeled Feature 1 and Feature 2; the 3D t-SNE coordinates are labeled Feature 3, 4, and 5. Default is disabled.

## PCA: Number of principal components

t-SNE uses principal components as its input. The number of principal components to use is set here.

We recommend using the PCA task to determine the optimal number of principal components for your data. Default is 50.

## PCA: Features contribute

Options are *equally* or by *variance*. Feature values can be standardized prior to PCA so that the contribution of each feature does not depend on its variance. To standardize, choose *equally*. To take variance into account and focus on the most variable features, choose *by variance*. Default is *by variance*.

## Normalization: Log transform data

You can choose to log transform the data prior to running PCA as part of t-SNE. Default is disabled.

## Normalization: Log base

If you are normalizing the data, choose a log base. Default is 2 when *Log transform data* is enabled.

## Normalization: Log offset

If you are normalizing the data, choose an offset. Default is 1 when *Log transform data* is enabled.

## References

[1] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.



Your Rating:  Results:  22 rates