# H-tSNE: Hierarchical Nonlinear Dimensionality Reduction.

Kevin C. VanHorn, Murat Can Çobanoğlu

**Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390**

**{kevin.vanhorn, murat.cobanoglu}@utsouthwestern.edu**

## Abstract

Dimensionality reduction (DR) is often integral when analyzing high-dimensional data across scientific, economic, and social networking applications. For data with a high order of complexity, nonlinear approaches are often needed to identify and represent the most important components. We propose a novel DR approach that can incorporate a known underlying hierarchy. Specifically, we extend the widely used t-Distributed Stochastic Neighbor Embedding technique (t-SNE) to include hierarchical information and demonstrate its use with known or unknown class labels. We term this approach "H-tSNE." Such a strategy can aid in discovering and understanding underlying patterns of a dataset that is heavily influenced by parent-child relationships. Without integrating information that is known *a priori*, we suggest that DR cannot function as effectively. In this regard, we argue for a DR approach that enables the user to incorporate known, relevant relationships even if their representation is weakly expressed in the dataset.

**Availability:** github.com/Cobanoglu-Lab/h-tSNE

## Introduction

Dimensionality reduction involves the elimination of features or random variables for a given dataset. DR is a crucial component in modern analysis techniques as the magnitude of data available to businesses, scientists, and public administration continues to grows rapidly [1]. This data can include text-based or multimedia content crucial to entertainment, research, and business sectors. The role of DR for feature selection and extraction and data preprocessing is especially pertinent when analyzing immense volumes of data with methods such as machine learning [2].

We focus on the use of dimensionality reduction for visualization. When high-dimensional data is difficult to visualize, DR can be effective for transforming the data to 2D or 3D for more interpretable results. This can be achieved through linear and nonlinear approaches. Nonlinear dimensionality reduction (NLDR) is preferable when capturing the local and global structure of the data, where linear DR tends to be faster and more effective for global patterns [2].

One of the most widely adopted and effective general-purpose NLDR approaches is t-SNE, a variation on Stochastic Neighbor Embedding [3]. We chose this as the foundation of our approach because of its ability to effectively represent real-world high-dimensional data. Our modification introduces a novel, general-purpose NLDR approach that incorporates an input hierarchy. Because our approach is visualization-centric, it can be manually weighted with a strength factor to better represent patterns in the data. Finally, we implement this technique for both known and predicted class-based hierarchies with the option to influence the strength of a predicted class' structural effect based on its likelihood.

## Related Work

Our solution aims to build upon t-SNE by including an input hierarchy and distancing factor. Formulated by Maaten and Hinton, t-SNE is currently the premier dimensionality reduction method used for high-dimensional data visualization [3]. This technique is a variation of Stochastic Neighbor Embedding, or SNE, which converts the Euclidian distance between high-dimensional data points into conditional probabilities that represent similarities. This variation adds a Student's t-distribution to SNE which solves the crowding problem that standard SNE suffers from.

Many techniques exist for dimensionality reduction but mainly act as a "black box." Examples of such methods include Sammon mapping, Curvilinear Components Analysis (CCA), Stochastic Neighbor Embedding (SNE), Isomap, Maximum Variance Unfolding (MVU), Locally Linear Embedding (LLE), and Laplacian Eigenmaps [7]. These are often effective with artificial data but struggle to maintain both local and global structure on real-world data.

To our knowledge, no modification has been introduced to enable a general-purpose NLDR technique to be influenced by a variable input hierarchy. Supervised and semi-supervised DR approaches exist but are not readily applicable to general-purpose visualization [5,6]. Similarly, techniques that integrate hierarchical information such as Hierarchical Manifold Learning are less concerned with visual analysis in lower dimensions [8]. Our work thus aims to remedy this black box effect that is often present when applying stochastic embedding approaches. We argue that if a known hierarchy exists for a dataset, it can provide benefit when integrated into the analysis. For example, such an approach can help to alleviate the case where important and explicit relationships in the data are ignored due to low expression.

## Methods

Following the formulation of t-SNE, we assume a given high dimensional input $\mathcal{X} = \{x_1, x_2, ..., x_N\}$. Prior to learning an $s$-dimensional embedding, the pairwise similarity between $x_i, x_j$ is defined by joint probabilities $p_{ij}$. t-SNE implements two symmetric conditional probabilities affected by a variable distance function $d(x_i, x_j)$.

$$p_{j|i} = \frac{exp(-d(x_i, x_j)^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-d(x_i, x_k)^2/2\sigma_i^2)}, \quad p_{i|i} = 0 \tag{1}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{2}$$

Our algorithm introduces an additional pairwise distance calculation $\theta_{ji}$ based on the input hierarchy and a following normalization. With a mean normalization, this results in the below modification to Equation 1:

$$p_{j|i} = \frac{exp((\mu - \theta_{ji} \cdot d(x_i, x_j)^2)/2\sigma_i^3)}{\sum_{k\neq i} exp((\mu - \theta_{ji} \cdot d(x_i, x_k)^2)/2\sigma_i^3)} \tag{3}$$

We found a min-max normalization to be more visually effective, tightening the graph by preventing outliers from drastically altering the scale of the lower-dimensional embedding (usually in 2D). With this alteration and the Euclidean distance function $d(x_i, x_j) = ||x_i - x_j||$, we have the finalized equation for $p_{j\,|\,i}$:

$$\frac{exp((min + \theta_{j\,i} \cdot ||x_i - x_j||^2)/2(max \cdot min)\sigma_i^2)}{\sum_{k\neq i} exp((min + \theta_{j\,i} \cdot ||x_i - x_k||^2)/2(max \cdot min)\sigma_i^2)} \tag{4}$$

We calculate an additional pairwise distance matrix using the hierarchy-informed function $\theta$. Given an undirected (not necessarily connected) graph $G$ with labels assigned to each node, we process input points $i, j$ in high-dimensional space with labels $i', j'$ respectively. We then find the shortest path distance of two undirected nodes $\overrightarrow{j'i'}$. If two nodes are in different connected components, then we return zero as the path length. Define $maxdist$ as the maximum length shortest path in $G$, i.e. $max(S)$ where $S$ is the set of possible shortest paths in $G$. Finally, let $str$ denote the strength of this function on the conditional probability $p_{j\,|\,i}$. Given that $str \in [0, 1]$ the function is then bounded by $[str, 1]$. In our implementation, the upper bound of this variable is unchecked and can be increased for a more dramatic effect. The resulting function is as follows:

$$\theta_{j\,|\,i} = (1 - str) + \frac{str(\overrightarrow{j'i'} - 1)}{maxdist - 1} \tag{5}$$

Using this formulation, a user can choose to input the adjacency matrix of the input graph $G$ and the strength of the hierarchical distancing factor. Graphs can be specified for either dimension of the input dataset. In the case that there is a known hierarchy for the dataset but labels are unknown, we additionally specify a feature-based method that additionally weights the strength of a predicted class' structural impact based on its likelihood. Given the probabilities $p_i$ and $p_j$ of labels $i'$ and $j'$ we define $m = min(p_i, p_j)$. We then weight the distancing factor $\theta$ with $m$ as follows:

$$\theta_{j\,|\,i} = (1 - m \cdot str) + \frac{m \cdot str(\overrightarrow{j'i'} - 1)}{maxdist - 1} \tag{6}$$

## Results

We analyzed the performance of our technique on the single-cell RNA-sequencing dataset specified in Zheng *et al.* [10]. We chose to target this high-dimensional dataset due to the ambiguity between T cell subclasses in the raw t-SNE plot. CD4+ and CD8+ T cells in [10] are highly interspersed making it difficult to discern detailed relationships between the subclasses. Additionally, there is a known hierarchy for the given classes, which motivated different approaches for a hierarchy-based model. In particular, our visualizations are produced from a subset of the filtered "Fresh 68k PBMCs (Donor A)" provided by 10x Genomics [11]. Prior to analysis, we normalize UMIs (unique molecular identifiers) by dividing by the total UMI counts in each cell. Following the UMI normalization process detailed in [10], we then multiply by the median across the sum of UMI counts for each cell, we finally perform a mean normalization on the resulting matrix. UMIs under a target variance are removed (we chose $\sigma^2 = 0.1$). Our implementation is based on Scikit-learn's 'TSNE' function [13] and Maklin's implementation [12].

The provided labels, bolded in Fig. 1, are derived from a gene-based hierarchy. Through a purely lineage-based method, CD34+ should be in a different connected component, but we chose to keep it as the head node of the tree due to its early role in hematopoiesis and sparsity in the dataset. Otherwise, parent-child relationships are upheld in the graph, with a dense set of nodes to distinguish between T cell types.

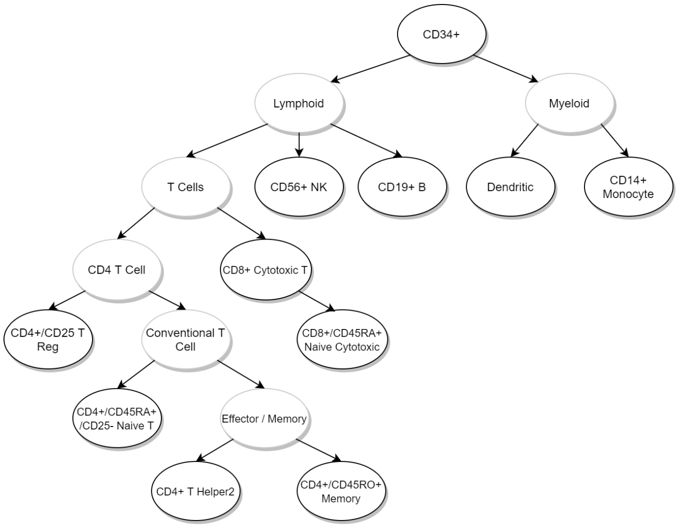

**Figure 1.** Class hierarchy for gene-based common lineages for the PBMC samples. Our hierarchical distance factor is based on the shortest path between two nodes in the graph. This clusters siblings and classes with a common lineage. If there is no path between nodes, the original pairwise distance is used.

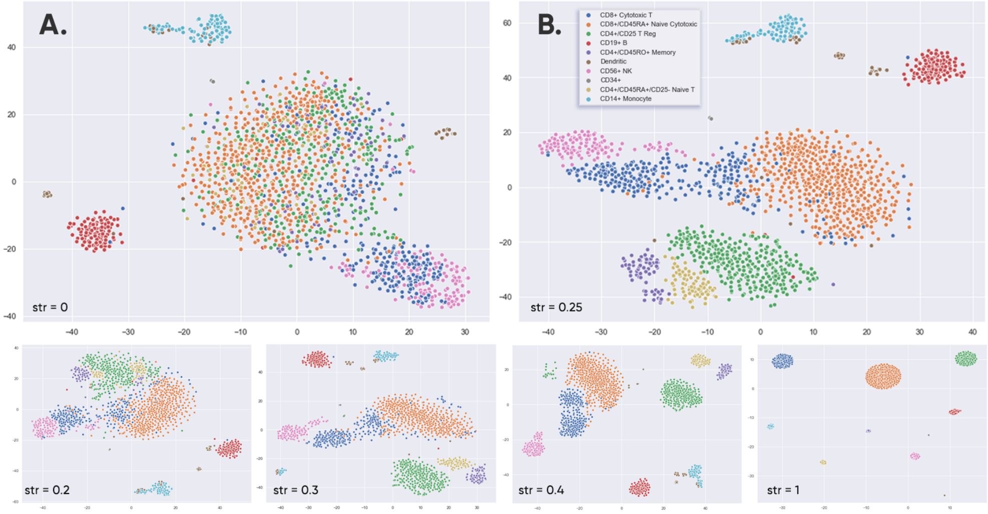

**Figure 2.** Visualization of h-tSNE for a filtered single-cell RNA-sequencing dataset for a cell-based hierarchy where labels are known. We display the progression of our approach at different distancing "strength" factors values. Compared to raw t-SNE (A), h-TSNE (B) is able to reveal sub-clusters and better express the relationship between datapoints.

We visualize our results in Fig. 2 using the aforementioned hierarchy. Here, the impact of the hierarchical distancing factor at different strengths is readily apparent in the progression of randomized embeddings. For this dataset, we favor a lower strength factor (0.25) to "push" clusters away from distant nodes in the hierarchy while maintaining the nuances of the raw t-SNE. In this regard, the variables not present in the hierarchy are still expressed in the higher dimensional manifold and visible within the embedding. Thus, the input graph and strength factor encourage a user-centric approach that allows the input of prior knowledge.

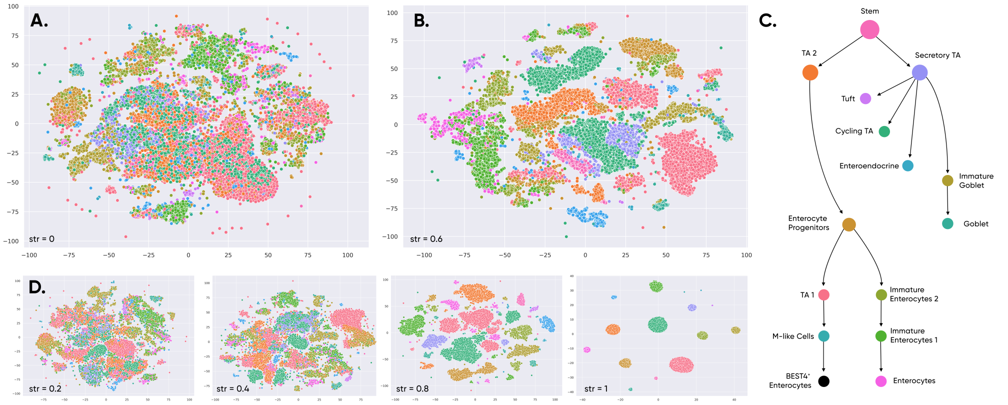

**Figure 3.** H-tSNE applied to an epithelial cell single-cell RNA-sequencing dataset [14]. Compared to raw t-SNE (A), h-TSNE (B) enables the identification of sub-clusters. At a strength factor of 0.6, Cycling TA separates into two distinct clusters (the same effect is seen with TA 1). At higher strength factors less than one (D), this separation is more apparent.

We additionally note a desirable divide between CD8+ and CD4+ T cells in Fig. 2(B). This separation still includes patterns that mirror the raw t-SNE plot (A) through the intermingling of NK and CD8+ Naive Cytotoxic classes. Original clusters of Monocyte, B, and CD34+ cells maintain relative distance and shape. In this manner, one can manually adjust the strength factor to set the degree of separation when targeting specific relationships in the graph.

This approach can similarly be performed on the same dataset for a gene-based hierarchy where sample classes are inferred with a given likelihood. H-tSNE implements a probability-dependent approach that weights poor classifications less heavily. Hierarchy nodes in this case can be derived from representative proteins of the provided classes. Using the filtered matrix normalized from $[0, 1]$, we can replace the ground-truth label of a sample $k$ with an inferred label $\ell'_k$ based on its total expression for a given node. Relevant UMI counts are first summed for each node $v \in V$ of the gene-based graph. We can then label the given sample based on the node with the highest normalized expression.

$$\ell'_k = max(\sum_{v \in V} \sum_{g \in v} M_{gk}) \tag{7}$$

Scores for each inferred class are then normalized from $[0, 1]$ and influence the likelihood-based pairwise distance matrix described in Section 3. Because the graph is also weighted with a strength factor, we can exaggerate this effect by increasing $str$. Inferred labels with low UMI counts contribute less to the hierarchical pairwise distance matrix, ensuring that the original embedding is upheld. Thus, the original structure can be preserved when the confidence of a given prediction is low. This prevents faulty clustering from the hierarchical pairwise distance matrix. With 100% confidence and 100% accuracy for inferred labels results of the gene-based method would mirror those of the cell-based hierarchy (Fig. 2).

To demonstrate the efficacy of our approach, we performed a similar cell-based process on a single-cell RNA-sequencing dataset comprised of epithelial cells [14]. We normalized UMI counts as before and filtered genes by a variance of 0.05. In Fig. 3, we indicate the effect of applying h-tSNE with the hierarchy (C) specified in [14]. Without our method, the raw t-SNE graph is too heterogeneous to discern any particular patterns in the data. Thus we integrate the known hierarchy to create a more readable embedding. For this dataset, we favor higher strength values due to the amount of UMI counts and higher number of classes. In Fig. 3(B), maintaining centralization of Cycling TA, TA 2, and Secretory TA cells within the graph. Most notably, by applying the hierarchy, we are able to observe sub-clusters that are not identifiable in the t-SNE graph.

## Discussion

H-tSNE formulates a direct relationship between the distance between two graph nodes in the hierarchy and the resulting distance in the embedding. We chose this approach to ensure that sibling nodes result a noticeable amount of spreading. The effect of the hierarchy can easily be modified for a desired effect on the embedding. To more strongly express a hierarchical effect, one could introduce a diminishing factor such that nodes deeper into the graph have a reduced influence as opposed to the base classes. An application of such a strategy is relevant with the presence of distinct classes containing many nuanced subclasses as children.

We chose to modify the pairwise distance function prior to embedding mostly due to the following random initialization, after which label ordering is unknown. However, with further modification of the base t-SNE algorithm, one could instead modify the distance function, influence the initial random embedding, or extend the KL-divergence objective function with a hierarchical factor. We also chose not to integrate directed graphs with this approach to support more complicated inter-class relations, but one could extend the approach to punish a pairwise-comparison when moving against the directed graph. This would result in more inter-class separation and less intra-class separation which could be valuable for some datasets.

We proposed a novel general-purpose approach for nonlinear dimensionality reduction that incorporates an input hierarchy. We modify the premier visualization-centric technique in this field, t-SNE, and demonstrate our results on a real-world single-cell dataset. Furthermore, we introduce a feature-based modification that enables users to integrate our method with weighted class labels.

## Acknowledgments

We thank Didem Çobanoğlu for her extensive guidance in the formulation of this approach and relevant biological expertise.

## References

1. U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” J. Bus. Res., vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.

2. “Dimensionality Reduction - an overview — ScienceDirect Topics.” https://www.sciencedirect.com/topics/computer-science/dimensionality-reduction (accessed May 02, 2020).

3. Laurens van der Maaten and Geoffery Hinton, ”Visualizing Data using t-SNE,” Journal of Machine Learning Research 9, 2579-2605, Nov. 2018.

4. Laurens van der Maaten and Eric Postma and Jaap van den Herik, "Dimensionality Reduction: A Comparative Review," 2009.

5. George Karypis and Eui-Hong Han, "Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization  Retrieval," Proceedings of the ninth international conference on Information and knowledge management, pp. 12-19, Nov. 2000.

6. Zhang, Daoqiang & Zhou, Zhi-Hua & Chen, Songcan, "Semi-Supervised Dimensionality Reduction," SIAM Data Mining, Apr. 2007, doi: 10.1137/1.9781611972771.73

7. T. Tr, "Dimensionality Reduction: A Comparative Review," p. 36.

8. Bhatia K.K., Rao A., Price A.N., Wolz R., Hajnal J., Rueckert D., "Hierarchical Manifold Learning," Medical Image Computing and Computer-Assisted Intervention, MICCAI 2012, Lecture Notes in Computer Science, vol 7510, 2012, doi: 10.1007/978-3-642-33415-3_63

9. V. B. Kolachalama and P. S. Garg, "Machine learning and medical education," Npj Digit. Med., vol. 1, no. 1, pp. 1–3, Sep. 2018, doi: 10.1038/s41746-018-0061-1.

10. G. X. Y. Zheng et al., "Massively parallel digital transcriptional profiling of single cells," Nat. Commun., vol. 8, no. 1, pp. 1–12, Jan. 2017, doi: 10.1038/ncomms14049.

11. "10x Genomics: Resolving Biology to Advance Human Health," 10x Genomics. https://www.10xgenomics.com/ (accessed May 02, 2020).

12. C. Maklin, "t-SNE Python Example," Medium, Aug. 12, 2019. https://towardsdatascience.com/t-sne-python-example-1ded9953f26 (accessed May 02, 2020).

13. "sklearn.manifold.TSNE—scikit-learn0.22.2documentation." https://scikit-learn.org/stable/modules/generated/sklearn. manifold.TSNE.html (accessed May 02, 2020).

14. C. S. Smillie, "Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis," p. 40.