High-Dimensional Data Visualization via Sampling-Based Approaches

Effect of Perplexity at different levels of Sampling-Based Approach

Muhammad Arslan Bhatti[1]

Supervisor(s): Klaus Hildebrandt[1]**, Martin Skrodzki**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Muhammad Arslan Bhatti
Final project course: CSE3000 Research Project
Thesis committee: Klaus Hildebrandt, Martin Skrodzki, Dr. C. (Christoph) Lofi

Abstract

Visualizing high-dimensional data is a key challenge in modern data analysis. T-distributed Stochastic Neighbor Embedding (t-SNE) is a popular nonlinear dimensionality reduction technique that maps such data into a low-dimensional embedding while preserving local relationships. A critical hyperparameter in t-SNE is *perplexity*. Choosing an appropriate value of perplexity for a particular use-case is non-trivial, especially for large datasets, where repeated t-SNE computations become computationally prohibitive. To mitigate this, the sample-based approach runs t-SNE twice: first on a downsampled subset of the data and then on the full dataset. This introduces two perplexity parameters: *sample perplexity* for the first run and *full perplexity* for the second run.

In this work, we systematically investigate the impact of varying combinations of sample perplexity and full perplexity on the quality of the final t-SNE embedding. Our findings show that sample perplexity predominantly determines the global layout of the embedding, while full perplexity influences local refinement. We also compare our approach with different approaches to choosing perplexity values, and find that while some offer better preservation of structural details, they provide less flexibility.

1 Introduction

Humans are inherently limited to visualizing and interpreting data up to a maximum of three dimensions. Yet, with the increasing collection of high-dimensional datasets in the modern era, particularly in fields like genomics and neuroscience, it is crucial to develop methods that make it possible to interpret high-dimensional data; for example, obtaining gene expression data from thousands or even millions of cells can help in identifying rare cell types or discover disease markers. [3; 19]. A promising technique introduced by [20], which is used to visualize high-dimensional data, is known as t-distributed Stochastic Neighbor Embedding (t-SNE); t-SNE aims to map high-dimensional data into a lower-dimensional embedding by preserving local similarities, such that points that are close to each other in the high-dimensional space are most likely to remain close in the embedding.

One of the most important parameters in t-SNE is perplexity, which determines how much emphasis is placed on nearby points when arranging the data, affecting whether the resulting visualization highlights local details or broader patterns in the data. Finding a suitable perplexity value for a given dataset is not straight forward because the optimal perplexity value depends on the structure of the dataset and also on the use-case of the person who wants to create the visualisations; it is a trial and error process of testing different perplexity values.

When t-SNE is applied to datasets exhibiting hierarchical structure, it presents two weaknesses: (i) t-SNE does not always preserve the global structure [21; 9] because it inherently preserves local neighborhoods; (ii) for very large datasets, the first weakness becomes worse [7], and running t-SNE multiple times to tune perplexity becomes computationally expensive [7].

To address these challenges, Kobak and Berens [7] propose the following techniques: PCA initialisation, use multi-scale similarities [9; 4], and increasing the learning rate parameter [2] in t-SNE; for very large data, in addition to the previously mentioned techniques, use sample-based approach and exaggeration. In the sample-based approach, t-SNE is run twice: first on a sampled subset and then on the full dataset. We refer to the perplexity used in the first step as the *sample perplexity*, and in the second step as the *full perplexity*. In [7], they use a combination of two values as the sample perplexity: 30 and $n/100$ where n is the number of data points, and use a fixed value of 30 as full perplexity.

As highlighted by a later study [17], the fixed sample perplexity used in [7] "restricts the user to a given, fixed perplexity". Building on the same sample-based strategy, [17] introduces flexibility in selecting the sample perplexity. It then adjusts the full perplexity "based on the sampling rate, leveraging the linear relationship between perplexity and dataset size" [17].

In our work, we further extend this approach [17] by making both sample perplexity and full perplexity tunable rather than be a fixed value or be heuristically determined, and aim to understand its effects on the quality of the final embedding. To this end, we conduct a grid search over a range of (sample perplexity, full perplexity) pairs and evaluate the resulting embeddings. Moreover, we then compare the outcome against the aforementioned approaches both qualitatively and quantitatively.

In this paper, we find that:

- Sample perplexity primarily determines the global structure of the embedding, serving as a foundation for the overall layout whereas full perplexity acts as a local refinement parameter where higher values smooth local neighborhoods but can reduce local details.
- Compared to the multi-scale approach by [7], our method significantly underperforms in preserving local structure.

2 Background

This section will contain the required knowledge and relevant concepts on which our research is based upon. In section 2.1 we will discuss the t-distributed Stochastic Neighbor embedding (t-SNE) algorithm. In section 2.2, we will discuss the role of hyperparameter: perplexity in the t-SNE algorithm. In section 2.3 we discuss the sampling-based approach in detail.

2.1 T-distributed Stochastic Neighbor Embedding (t-SNE)

The T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm is a popular method for visualizing high-dimensional data in two or three dimensions. It operates by minimizing the Kullback-Leibler (KL) divergence between

two probability distributions: one defined over pairs of points in the original high-dimensional space, and the other over the corresponding points in the low-dimensional embedding space [20].

In high dimensions, the similarity between points is modeled using a Gaussian distribution. The similarity between two points x_i and x_j in the high-dimensional space is computed using

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

where $p_{i|i} = 0$ and σ_i^2 is the variance of the Gaussian centered on the point x_i.

These conditional probabilities are then symmetrized to form joint probabilities using

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

where n is the total number of data points in the dataset.

In the low-dimensional space, the similarity between points y_i and y_j is modeled using a Student-t distribution as

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}}. \tag{3}$$

In t-SNE, an initial embedding is obtained by principal component Analysis (PCA) to compute the positions y_i of the low-dimensional embedding [8]; this low-dimensional embedding is then altered by gradient-descent optimization of the Kullback-Leibler (KL) divergence between the high-dimensional and low-dimensional similarity distributions, P and Q given by

$$\mathrm{KL}(P\|Q) = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{4}$$

with the gradient

$$\frac{\partial \mathrm{KL}(P\|Q)}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})\left(1 + \|y_i - y_j\|^2\right)^{-1}(y_i - y_j). \tag{5}$$

2.2 Hyperparameter: Perplexity

A key hyperparameter in t-SNE is *perplexity*, which determines the effective number of neighbors considered for each point during the computation of similarities. Formally, for each data point x_i, t-SNE fits a Gaussian distribution centered at x_i and adjusts its variance σ_i^2 such that the *Shannon entropy* of the resulting conditional similarity distribution $P_i = \{p_{j|i} \mid j \in [n]\}$ matches a target perplexity value. This is done via binary search on σ_i [20] to solve:

$$\text{Perplexity}(P_i) = 2^{H(P_i)} \tag{6}$$

where H is the Shannon entropy of a discrete distribution given by:

$$H(P_i) = -\sum_j p_{j|i}\log_2 p_{j|i}. \tag{7}$$

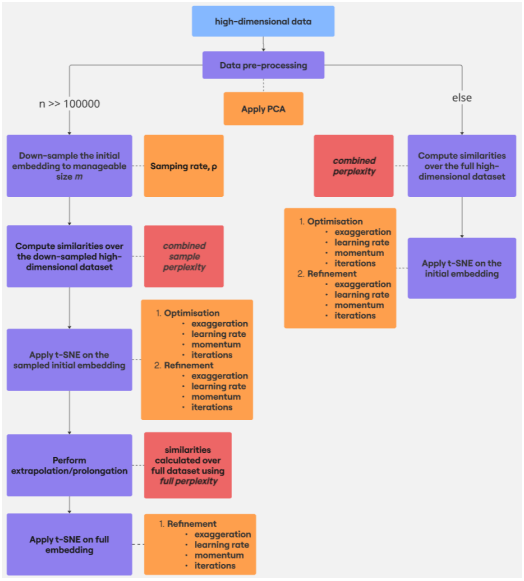


Figure 1: Kobak and Berens approach for the standard case and very large data sets

As a result, lower perplexities, w.r.t. the dataset size lead to smaller σ_i^2, producing narrow Gaussian that emphasize nearby points yielding embeddings that capture *fine-grained, local structure*. Conversely, higher perplexities, w.r.t. the dataset size increase σ_i^2 spreading the Gaussian wider and incorporating more distant neighbors which helps emphasize *broader, global relationships* among clusters [7]. Hence, perplexity acts as a smooth control over the local versus global tradeoff in the structure of the embedding.

2.3 Sampling-Based approach

Selecting an perplexity value which is suitable for a particular use-case is non-trivial and often dataset dependent. Evaluating multiple perplexity values to explore the embedding space becomes computationally expensive, particularly for large datasets, due to t-SNE quadratic complexity in both time and memory [20]. A sample-based approach introduced by [7] mitigates the issues which employs the following pipeline:

(i) Downsample the dataset to a manageable size.

(ii) Apply t-SNE on the sample using a specified *sample perplexity*.

(iii) Position all the remaining unsampled data points on the resulting t-SNE embedding using nearest neighbours.

(iv) Use the result as initialization to run t-SNE on the whole dataset using a specified *full perplexity*.

Kobak and Berens approach

Kobak and Berens [7] proposed several practical improvements to classical t-SNE with a focus on preserving global structure in single-cell transcriptomic datasets which is hierarchical in nature. However, these improvements are applicable to any dataset with hierarchical structure [16]. [7] introduces three

key improvements: (i) multi-scale similarities [9; 4], (ii) PCA initialisation, (iii) high learning-rate [2].

This work highlights the issue that "as large perplexity yields longer-ranging attractive forces during t-SNE optimisation, the visualisation loses some fine detail but pulls larger structures together" [7].. To curb this issue, this work proposes a multi-scale approach [9; 4]. Specifically, for each data point i, similarities to other points are computed using two Gaussian kernels with different variances, corresponding to two perplexities e.g. 30 and $n/100$. The resulting multi-scale kernel is defined as

$$\frac{1}{\sigma_i}\exp\left(-\frac{d^2}{2\sigma_i^2}\right)+\frac{1}{\tau_i}\exp\left(-\frac{d^2}{2\tau_i^2}\right) \tag{8}$$

where d is is the Euclidean distance between points, and σ_i^2, τ_i^2 are selected such that the perplexity of the first Gaussian corresponds to the lower scale e.g. 30, and the second to the higher scale e.g. $n/100$. This formulation effectively averages over neighborhood sizes at different scales.

[7] also includes the use of PCA initialisation and a higher learning rate. PCA initialisation "injects the global structure into the t-SNE embedding which is then preserved during the course of t-SNE optimisation" [7]. Moreover, a high learning rate helps preserve the local structure by allowing the optimisation to escape poor local minima and converge more quickly to a configuration that better reflects neighborhood relationships; it uses the learning rate $\eta = n/12$ as suggested by [2].

For very large datasets ($n >> 100,000$), in addition to the above recommendations, Kobak and Berens uses exaggeration and the sampling-based approach. Exaggeration is applied at two key steps: in Step (ii), a high exaggeration coefficient (typically 12 [13]) is used; in Step (iv), a constant exaggeration factor (between 1 and 10 [11]) is applied when refining the full dataset embedding.

Figure 1 provides an overview of the steps involved in[7], illustrating both the case of very large datasets and the standard case. The purple blocks gives the description of each step; the red blocks represent the perplexities used at different steps, and the orange blocks represent the method and the parameters used in that step.

3 Related Work

Similar to [7], there exist several recent works which address the challenge of applying t-SNE to large datasets, and the choice of perplexity value. In this section, we first review techniques designed to reduce the computational cost of t-SNE (Section 3.1), followed by methods that explore strategies for selecting perplexity values (Section 3.2).

3.1 Accelerating t-SNE

One of the techniques to accelerate t-SNE is FIt-SNE [12]. It addresses the computational cost involved in estimating gradients, especially due to the quadratic complexity of computing repulsive forces between all pairs of points; it does so by interpolating the expensive repulsive-force computation onto an equispaced grid and using the Fast Fourier Transform on this grid to speed up convolution, which approximates the repulsive forces. This makes gradient descent orders of magnitude faster for millions of points. FIt-SNE also proposes a late-exaggeration variant and an out-of-core PCA implementation to sharpen cluster boundaries and enable embedding of data sets that do not fit in memory. FIt-SNE also accelerates the computation of high-dimensional input similarities by using multi-threaded approximate nearest neighbor search.

Another related technique to accelerate t-SNE is opt-SNE [2]; this technique focuses on automating key parameter choices in t-SNE to improve its performance on large datasets. Essentially, opt-SNE monitors the Kullback Leibler divergence (equation 4) in real time to tailor early-exaggeration, learning rate and the number of iterations in the gradient-descent. Combined with a more adaptive iteration schedule and multi-threaded implementation, opt-SNE produces high-quality visualizations with fewer iterations and less manual tuning. These optimisations overcomes the limitations of hard-coded parameters which produce low quality and misleading embeddings of mass cytometry data.

3.2 Choice of perplexity

Other related work focus on the choice of perplexity rather than speed. One recent work [17] follows the sampling-based approach mentioned in section 2.3, and proposes that "embedding remain structurally similar when scaling perplexity linearly with sample sizes"' [17] which preserves the embedding's key visual characteristics across different sample sizes; this proposition is motivated by the intuition that "the optimal perplexity parameter increased as the total number of data points increased" [5]. Formally,

$$\textit{full perplexity} = \frac{\textit{sample perplexity}}{\rho} \tag{9}$$

where ρ is the sampling rate.

This approach [17], unlike [7], makes the sample perplexity tunable for the user while deriving the full perplexity using a linear scaling rule (Equation 9). However, this still limits the exploration of how different combinations of sample and full perplexities affect the embedding quality.

In our work, we generalize the flexibility of the t-SNE framework by making both sample perplexity and full perplexity independently tunable. This allows us to explore the extent to which the change in the sample perplexity and full perplexity has on the resulting embeddings more deeply which has not been explored in the aforementioned work. Moreover, we thoroughly compare quality and structure of the embeddings with the previous approaches [7; 17]. Table 1 gives an overview of the choice of perplexity for each approach, highlighting the unique flexibility and experimental scope of our work.

4 Methodology

In Section 4.1, we mention the methodology used to analyse our approach. Section 4.2 outlines the methodology used to compare the different approaches.

Dataset Size	Kobak and Berens	Linear scale	Our approach
$n \gg 100{,}000$	combined sample perplexity = 30 & $\frac{n}{100}$ full perplexity = 30	sample perplexity = s full perplexity = s/ρ	sample perplexity = s full perplexity = f
$n < 100{,}000$	combined perplexity = 30 & $\frac{n}{100}$	sample perplexity = s full perplexity = s/ρ	sample perplexity = s full perplexity = f

Table 1: Comparison of sample and full perplexity settings across different approaches. Here, n denotes the number of data points, s and f represent the sample and full perplexity values, respectively, which are tunable. In the linear scale method, ρ is the sampling rate used to derive full perplexity from the sample perplexity.

4.1 Analysis of our approach

To systematically investigate the effect of varying sample and full perplexity values, we perform a grid search across a defined range of (sample perplexity, full perplexity) pairs. For each pair, we compute the corresponding t-SNE embedding.

Then these embeddings are qualitatively analysed by visual inspection of their cluster structure. Specifically, we assess how well clusters are separated, whether any fragmentation occurs within clusters; we also analyse structure preservation i.e. how well are the local and global structures preserved, and also consider the overall interpretability of the embedding layout. We also analyse quantitatively using the following metrics:

- **KNN (k-Nearest Neighbor):** Measures local structure preservation by computing the fraction of each point's k-nearest neighbors in the high-dimensional space that remain among its k-nearest neighbors after embedding [10]. We use $k = 10$ and average over all n points.
- **KNC (k-Nearest Class Means):** Evaluates preservation of inter-class relationships by checking whether each class's k-nearest class means in the original space remain the same in the embedding. Captures mesoscopic structure; we set $k = 4$.
- **CPD (Cross-Pairwise Distance Correlation):** Quantifies global structure preservation via Spearman rank correlation between pairwise distances in the original and embedded spaces [1]. Computed on 499,500 pairs from 1,000 randomly selected points.

4.2 Comparing different approaches

To compare the resulting embeddings from our approach with Kobak and Berens approach, we use data sets of varying sizes, including those with fewer than 100,000 points and those with significantly more since [7] uses different perplexity settings for different dataset sizes as mentioned in Table 1. Most importantly, we replicate the perplexity settings in their work while keeping all other parameters fixed to ensure a fair evaluation of the effect of perplexity.

A core challenge is comparing our grid of embeddings with a single embedding produced by [7]. Besides visual inspection, to enable a controlled comparison, we manually generate embeddings using our approach by setting the sample and full perplexities to match the smaller and larger perplexity values, respectively, used in their combined perplexity. This enables a direct comparison of whether combined or separate handling of perplexities leads to better structural preservation and embedding quality.

Finally, to incorporate comparisons with [17], for each sample perplexity, we select full perplexity values that follows the scaling (equation 9).

5 Experimental Setup

The experimental setup used is closely related to [17]. For all the experiments mentioned, we will be using the FIt-SNE implementation of t-SNE [12]. Prior to applying t-SNE, we perform a preprocessing step for each dataset. If the input dimensionality exceeds 50, we first reduce it to 50 using Principal Component Analysis (PCA) [8].

Unless explicitly stated otherwise, all experiments use the following t-SNE parameters: the first optimisation stage begins with an early exaggeration phase, where gradient descent is run for 250 iterations using an exaggeration factor of 12 [2], a momentum of 0.5 [6; 20], and a learning rate of $\eta = n/12$ as suggested by [2]; this is followed by the normal regime phase, which continues for 750 iterations with the exaggeration factor reduced to 1 and the momentum increased to 0.8 [6; 20]. In the case of sampling-based approach, a second optimisation stage is required which follows the prolongation step; it runs for 750 iterations with an exaggeration factor of 1 and momentum set to 0.5. Moreover, we will use the sampling rate $\rho = 0.1$, and a fixed seed of value 42 to ensure the reproducibility of the results. To make sure that the choice of seed does not affect the outcome of our experiments, we tested different seed values. Hence, for every combination of sample perplexity and full perplexity, the same subset of points are sampled to ensure fair comparison.

Datasets

It is important to select the correct datasets which cover both cases of [7] as shown in Figure 1, and also exhibit Hierarchical structure to examine the property of local and global structure preservation. To this end, we use the following datasets: MNIST dataset is a widely used dataset to test dimensionality reduction algorithms. The Tasic et al. 2018 dataset [19] and the Wong dataset are both single-cell transcriptomic datasets from the mouse brain, but they differ in scope, resolution, and biological context. Tasic et al. dataset is widely used to benchmark clustering, visualization, and embedding algorithms due to its hierarchical structure

sample perplexity	full perplexities MNIST					full perplexities Tasic et al.				full perplexities Wong				
10	10	30	300	500	1000	10	30	100	300	10	30	100	300	500
30	10	30	300	500	1000	10	30	100	300	10	30	100	300	500
100	-	-	-	-	-	10	30	100	300	10	30	100	300	500
300	10	30	300	500	1000	10	30	100	300	10	30	100	300	500
500	10	30	300	500	1000	-	-	-	-	10	30	100	300	500
1000	10	30	300	500	1000	-	-	-	-	N/A	N/A	N/A	N/A	N/A

Table 2: Sample perplexity and full perplexity values used for the embeddings in Figure 7. N/A denotes combinations which could not be supported by the hardware used for these experiments.

and fine-grained annotations; Wong dataset has a broader resolution, and is a relatively large dataset. A summary of the data sets used and their properties can be found in Table 3.

Name	Data Type	# Pts.	# Dim.	# Cl.
MNIST	images	70,000	784	10
Tasic et al.	single-cell	23,822	30000	133
Wong	single-cell	372,674	11	6

Table 3: Data sets used for experiments in Section 5. For each data set, the size, dimension of data points and the number of classes are given.

Perplexity values

It is important to mention the choice of perplexity values we have used for the grid. To simulate the tuning of sample perplexity and full perplexity and analyse the extent to which the change in these values affect the embeddings, we will choose the maximum value our hardware could support. Since "perplexity values in the common range (e.g. 20, 50, 80) yield similar results" [7], we will choose values which are not in the common range. Table 2 gives a summary of the sample perplexity and full perplexity used in our experiment. Lastly, to make sure that we also plot the embeddings from the approach mentioned in [17], we use combinations such as (10, 100) and (30, 300) with the sampling rate $\rho = 0.1$.

6 Results

In the Section, we discuss the results from the experiments in detail. In Section 6.1, we will analyse both qualitatively and quantitatively. In Section 6.2, we will compare the approaches.

6.1 Analysis of our approach

We made several key observations regarding the visual appearance of the embeddings in the grid (Figure 7). For a fixed sample perplexity, increasing the full perplexity does not substantially alter the overall visual structure of the embeddings. The clusters retain a similar shape and arrangement. At the same time we notice that as the full perplexity increases, the clusters begin to overlap more; for example, in the second row of the MNIST dataset, the blue and purple clusters are highlighted across the row to visualise how the structures remain visually very similar. Specifically, highlight 1 and 2 (4a), and highlight 5 and 6 (Figure 4b) show the overlapping of the clusters with increasing full perplexity for the MNIST dataset and Tasic et al. dataset respectively.

Dataset	Sample Perp	Full Perp	CPD
MNIST	30	30	0.990
	30	500	0.959
	30	1000	0.954
Tasic et al.	30	30	0.992
	30	100	0.977
	30	300	0.939

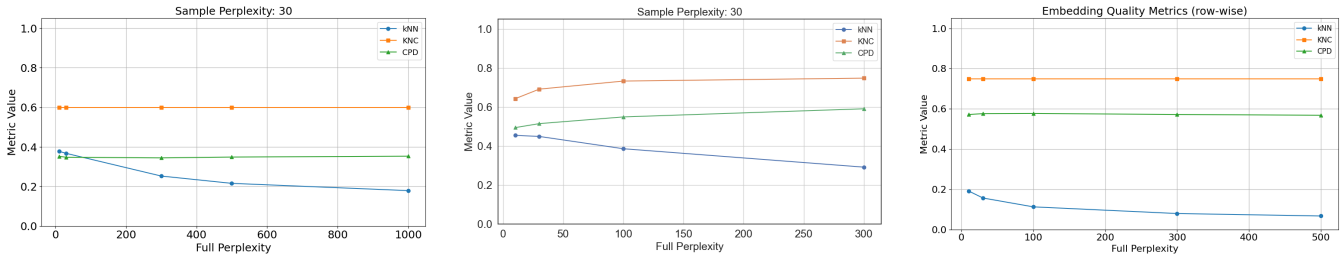
Table 4: CPD values calculated to quantify similarity between embeddings in a row for the MNIST and Tasic et al. dataset, where the reference embedding for both dataset has sample perplexity 30 and full perplexity 30.

Dataset	Sample Perp	Full Perp	CPD
MNIST	300	500	0.861
	500	500	0.782
	1000	500	0.643
Tasic et al.	30	100	0.920
	100	100	0.397
	300	100	0.546

Table 5: CPD values calculated to quantify similarity between embeddings in a column for the MNIST and Tasic et al. dataset, where the reference embedding for the MNIST dataset has has sample perplexity 10 and full perplexity 500, and the Tasic et al. dataset has has sample perplexity 10 and full perplexity 100

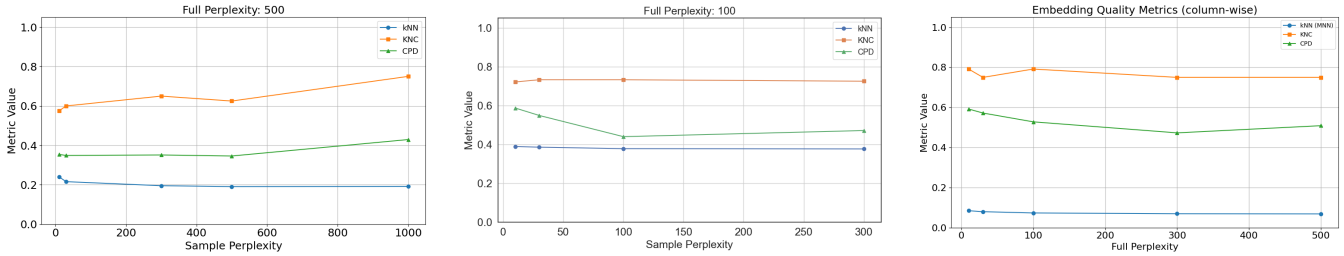
The trends mentioned are corroborated by quantitative metrics as well. We compute the CPD between the first embedding in each row and all subsequent embeddings in the same row to quantify the structural similarity across different full perplexity values; in Table 4, we see very similar values for same sample perplexity value and different full perplexity values. Moreover, the overlapping of clusters with increasing full perplexity suggest loss of local details [7]; when plotting the KNN for embeddings across the row, we observe a decline in KNN value across all datasets in Figure 2.

In contrast, for a fixed full perplexity, varying the sample perplexity does not keep the visual structure of the embeddings similar. This is evident from the embeddings of the MNIST dataset (Figure 4a) where in highlight 3, the brown cluster is not fragmented; however, in highlight 4, the brown cluster is fragmented into 3 separate parts. Table 5 shows the CPD values calculated across the column where the reference embedding is the first embedding in the column; the table shows different values for each embedding which confirms that embeddings in a column are not structurally similar. Moreover, we also observe that, unlike fixed sample perplexity, the arrangement of the clusters can change



(a) MNIST (b) Tasic et al. (c) Wong: Sample perplexity 30

Figure 2: Quality metrics (KNN, KNC, CPD) plotted for each dataset for fixed sample perplexity and varying full perplexity. For all three datasets, sample perplexity is 30.



(a) MNIST (b) Tasic et al. (c) Wong: full perp 300

Figure 3: Quality metrics (KNN, KNC, CPD) plotted for each dataset for fixed full perplexity and varying sample perplexity. The full perplexity values are 500, 100, 300 for MNIST, Tasic et al., and Wong dataset respectively.

significantly; for example, in the embedding of the Tasic et al. dataset (Figure 4b), in highlight 7, the red cluster is above the yellow cluster, but in highlight 8, the red cluster is on the left side of cluster 8. This signifies that the local structure stays the same, but the global structure can vary; this observation is also evident in the graphs in Figure 3 in which the KNN value remains relatively constant, but the CPD (between the high-dimensional data and the low-dimensional embedding) and the KNC value vary.

It is evident from the results that sample perplexity is responsible for defining the global structure of the embeddding and full perplexity whereas full perplexity is responsible for applying local refinements.

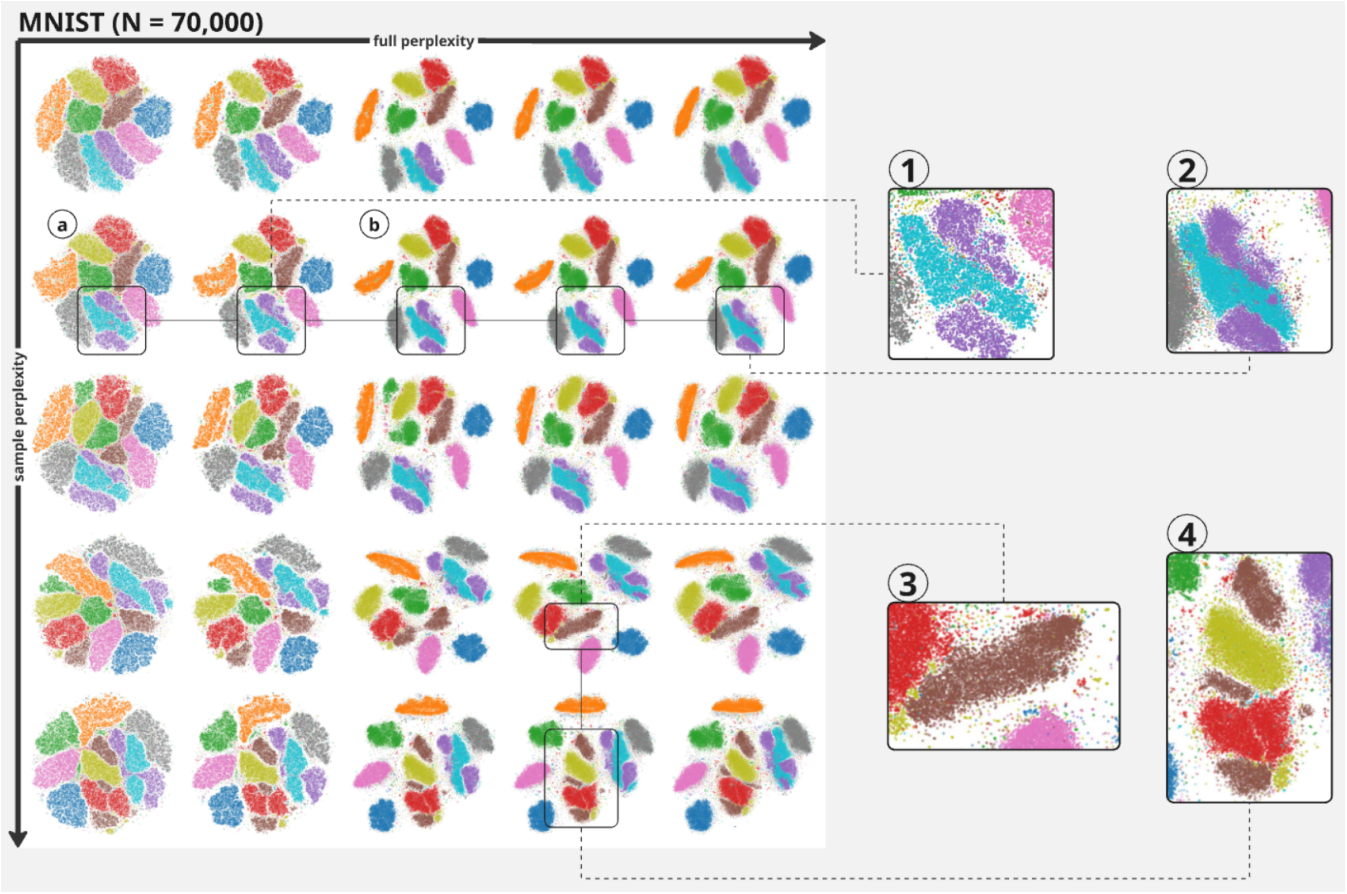
6.2 Comparing different approaches

Following the methodology described in Section 4.2, we align the perplexity configurations of both approaches by setting our sample perplexity to the lower perplexity used in [7], and the full perplexity to the higher one.

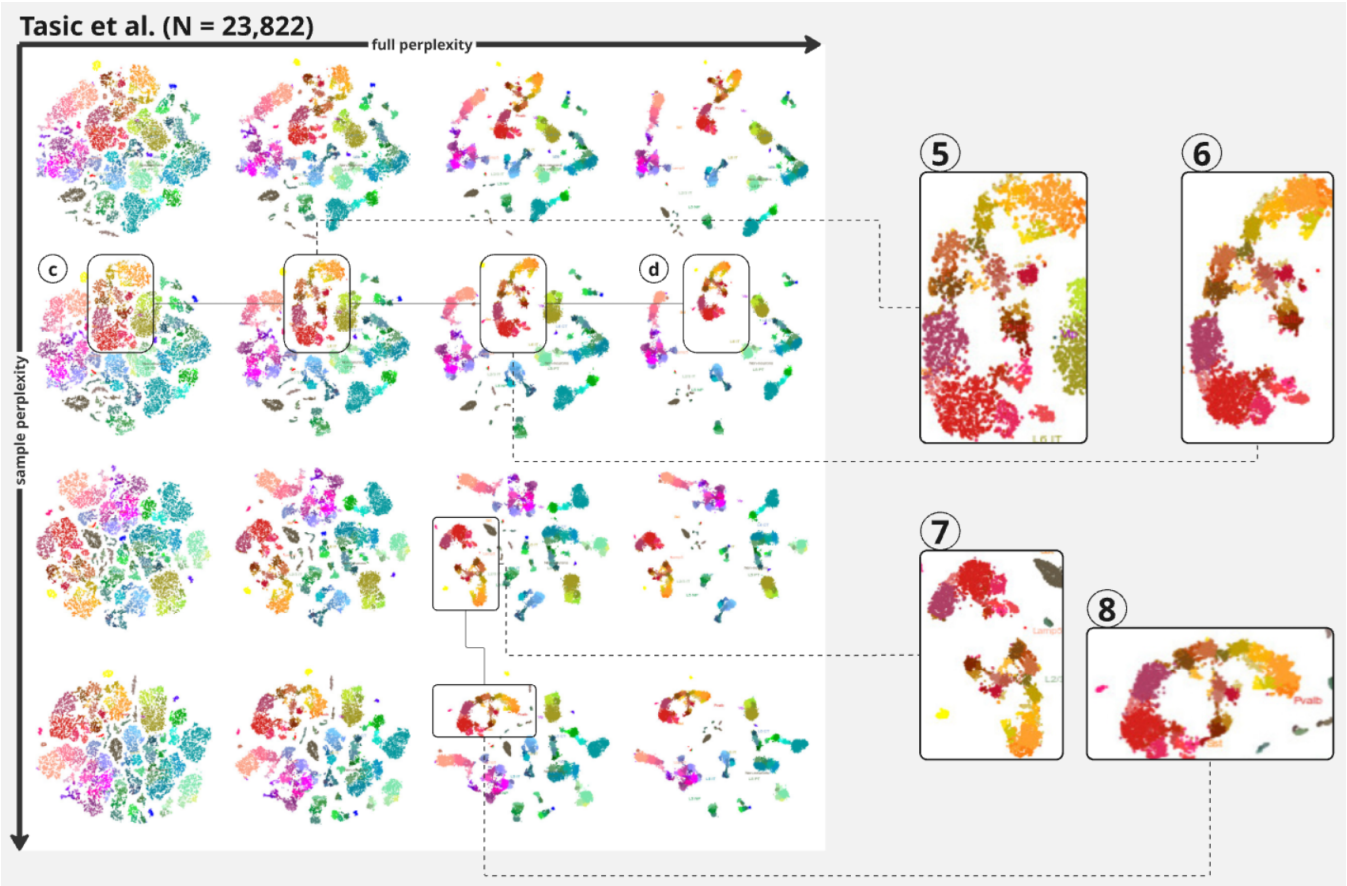
We first compare the embeddings produced for MNIST dataset by both approaches. [7] uses a combined perplexity of 30 and 700; hence, for our approach we set the sample perplexity to 30 and full perplexity to 700. The MNIST embedding by [7] (Figure 6a) exhibits distinct and well-separated clusters with enough white space between them to be distinguishable. Although the blue and purple clusters overlap creating fragmentation, there is enough white space to be distinguishable compared to the embedding produced by our approach (Figure 5a) in which the blue and purple clusters overlap more hence losing the local refinements; this is also evident from the lower KNN value of the embedding produced by our approach than the other approach. Our embedding also has a lower CPD value compared to [7] embedding which means that [7] also preserves global structure better.

Next, we compare the Tasic et al. embeddings; this dataset [19] has three well-separated group of clusters: excitatory neurons (cold colors), inhibitory neurons (warm colors), and non-neural cells (grey/brown colors). Embedding from our approach (Figure 5b) and [7] (Figure 6b), both preserve the global structure well where the three group of clusters are well-separated; this is also evident from the similar value of the CPD metric. However, our embedding has clusters which start to overlap losing the local refinements hence a lower KNN value. Similarly, the embeddings for the Wong dataset from our approach (Figure 5c) exhibits more fragmentation in the green cluster than the other approach [7] hence a lower KNN value.

To compare our grid of embeddings with the embedding produced by [7], we adopt a structured two-step methodology based on our observations from Section 6.1: sample perplexity primarily determines the global layout, while full perplexity acts as a local refinement mechanism. Accordingly, we first identify the embedding in our grid, specifically from the first column of Figure 7, where full perplexity is minimal, whose global structure most closely resembles that of the embedding by [7]. This step ensures a fair comparison by aligning on global structure, which in our method is governed by sample perplexity. Once the global layout is matched, we analyse how the change in full

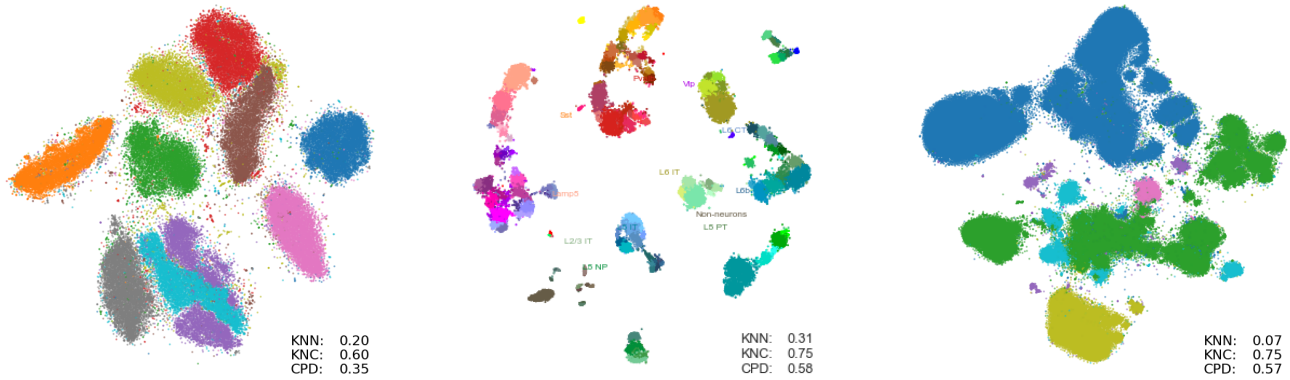


(a) Embeddings of the MNIST dataset



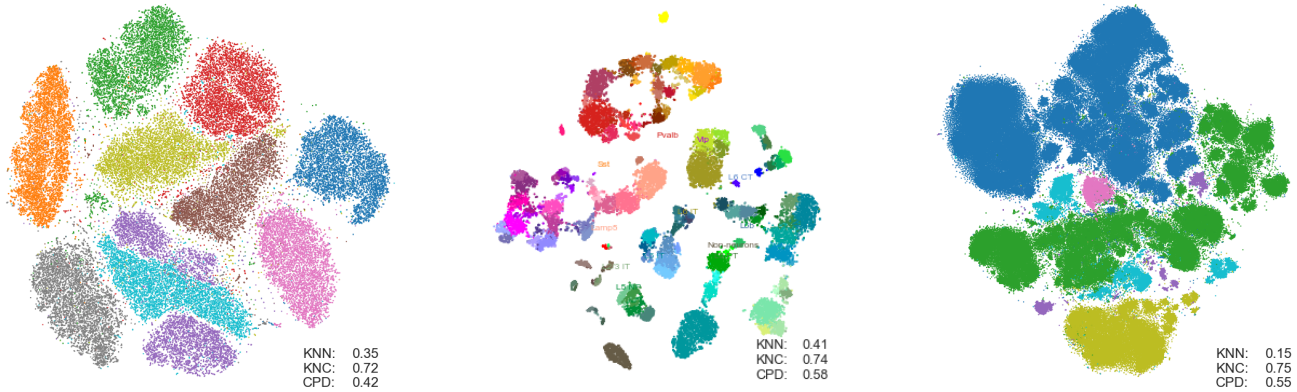
(b) Embeddings of the Tasic et al. dataset

Figure 4: Embeddings resulting from our approach on MNIST (a) and Tasic et al. (b).



(a) MNIST (b) Tasic (c) Wong

Figure 5: Embeddings produced by our approach. Metrics shown: KNN, KNC, and CPD.



(a) MNIST (b) Tasic (c) Wong

Figure 6: Embeddings produced by the method of Kobak and Berens [7]. Metrics shown: KNN, KNC, and CPD.

perplexity modifies this layout. This allows us to isolate and compare the refinement behavior of our method against that of [7] ensuring that the observed differences are attributed to local refinement mechanisms rather than differences in global layout.

For instance, in the MNIST embeddings (Figure 4a), the closest embedding visually matching the global layout of Figure 6a is highlighted **a** which has the sample perplexity 30. We observe that as full perplexity increases, the KNN value declines. Following the approach of [17], if we scale the sample perplexity accordingly, we get the embeddings highlighted as **b**; this embedding has a KNN value of 0.25, notably lower than that of [7]. We observe a similar trend in the Tasic et al. embeddings, confirming that our method exhibits a decline in local neighborhood preservation with the increase in full perplexity than that of [7].

7 Responsible Research

Responsible research is important to ensure that scientific advancements are developed in ethical and trustworthy manner. We discuss the steps we have taken to ensure reproducibility in 7.1. The ethical considerations that should be taken into account are mentioned in Section 7.2. Finally, we discuss the use of LLM tools for this research in Section 7.3.

7.1 Reproducibility

Reproducibility is important for another group of researchers to achieve the same result with the original data if they want to expand on this work or get a better understanding of the topic, enabling both verification and further exploration. For this paper, we have taken the required steps for not only transparency but also reproducibility for the future as mentioned in [18]. The code used in this research paper can be found in https://github.com/mearslanbhatti/sample-based-tsne, and a digital copy of this paper is also available at respository.tudelft.nl.

Our research builds upon the prior open-source work developed in previous researches; we have mentioned the open source repositories and also the corresponding papers in the README file of our github repository to ensure proper credit and transparency. Moreover, the repository also contains datasets used in the experiments.

The datasets used in this study are all publicly available and also widely used in the research community which supports

reproducibility and transparency. The MNIST dataset (https://yann.lecun.com/exdb/mnist/) is available under the Creative Commons Attribution-Share Alike 3.0 license. The Tasic et al. dataset (https://portal.brain-map.org/atlases-and-data/rnaseq) is publicly released by the Allen Institute for Brain Science under their open data sharing policy. Lastly, the Wong dataset (http://flowrepository.org/id/FR-FCM-ZZTM) is similarly available through public repositories.

7.2 Ethical considerations

In our research, we have used datasets such as Tasic et al. [19] (single-cell transcriptomics from mouse cortex) and the Wong dataset (neuronal recordings). Techniques such as t-SNE are powerful tools to visualise high-dimensional data, but if the original properties of this high-dimensional data are not properly preserved in the lower dimensional embeddings, such visualisations may lead to misleading biological conclusions for instance in the case of datasets like Tasic et al., these distortions could influence the classification of neuronal subtypes, which in turn affects our understanding of brain function or disease. As emphasized in [14; 15], visualisations can be persuasive, even when their underlying assumptions are misunderstood. Hence, the plots from our research should be regarded as exploratory and not definitive. Any interpretations or conclusions drawn from these plots should be supported by domain expertise and additional validation.

7.3 Use of LLM tools

Large Language Models (LLMs), particularly ChatGPT, were used throughout the research process to support and enhance understanding. Below we mention how it was used.

- Initially, it was used to understand the algorithms introduced in previous work. The tool was always used as a secondary source to enhance the understanding while keeping the original work as the primary source of understanding.
- Additionally, the tool was used to rephrase sentences for improved clarity and articulation. We strictly refrained from including any large body of texts generated from LLM. It was also used to edit the equations for latex format. It was also used to generate latex format tables.
- It has also been used in helping to write and understand existing code; However, it was not used to generate large portions of code, as the research heavily relies on existing implementations. Its usage was limited to writing small code snippets within methods or exploring alternative approaches.
- It was used as a brainstorming partner for idea generation, but all key decisions, implementations, and interpretations were made independently by us after thorough evaluation.

A list of key prompts can be found in the Appendix that demonstrates the use of LLM tools in our research.

8 Conclusion and Future Work

In this paper, we investigate a sample-based approach of running t-SNE in which the algorithm is run twice, first on the downsampled subset and then on the full subset, introducing two tunable perplexity parameters: sample perplexity and full perplexity. Unlike previous approaches that fix one of these values or derive it heuristically, our approach treats both as independently tunable. To investigate the affect of these perplexity values on the final embedding, we generate a grid of embeddings across a range of (sample perplexity, full perplexity) pairs.

Our results show that sample perplexity is the dominant factor in determining the global structure of the embedding: different sample perplexity values with fixed full perplexity generate very different cluster layouts and inter-cluster relationships as well. In contrast, once the global structure is established, full perplexity primarily refines local structure. We found that although increasing full perplexity smooths neighborhood relationships, it can also lead to cluster overlap and hence a loss of local structure which is also evident from declining KNN values.

We also compared our approach with the multi-scale approach of [7] in which we find that the multi-scale approach achieved higher local structure preservation, as measured by KNN, consistently across all tested datasets than any full perplexity setting in our grid which shows its strength of capturing fine-grained structure. Similarly, the multi-scale approach also performed better in preserving the local structure than the linear scaling heuristic proposed in the Navigating paper [17], which ties full perplexity to sample size.

Based on our findings, we recommend using our approach in scenarios where interpretability and user control are important, as it offers greater flexibility in shaping the embedding through independent tuning of sample and full perplexities. However, the user still has to select the values manually. For future work, a strategy can be developed to automatically suggest suitable values based on the use case and dataset characteristics making our approach more accessible.

References

[1] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[2] Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature communications*, 10(1):5415, 2019.

[3] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[4] Cyril De Bodt, Dounia Mulders, Michel Verleysen, and John A Lee. Perplexity-free t-sne and twice student tt-sne. In *ESANN*, pages 123–128, 2018.

[5] Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications*, 9(1):2002, 2018.

[6] Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

[7] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature Communications*, 10(1):5416, 2019.

[8] Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature biotechnology*, 39(2):156–157, 2021.

[9] John A Lee, Diego H Peluffo-Ordóñez, and Michel Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[10] John A Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431–1443, 2009.

[11] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Efficient algorithms for t-distributed stochastic neighborhood embedding. *arXiv preprint arXiv:1712.09005*, 2017.

[12] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245, 2019.

[13] George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM journal on mathematics of data science*, 1(2):313–332, 2019.

[14] Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. Misinformed by visualization: What do we learn from misinformative visualizations? In *Computer Graphics Forum*, volume 41, pages 515–525. Wiley Online Library, 2022.

[15] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. The persuasive power of data visualization. *IEEE transactions on visualization and computer graphics*, 20(12):2211–2220, 2014.

[16] Benjamin Schmidt. Stable random projection: Lightweight, general-purpose dimensionality reduction for digitized libraries. 2018.

[17] Martin Skrodzki, Nicolas F. Chaves de Plaza, Thomas Höllt, Elmar Eisemann, and Klaus Hildebrandt. Navigating perplexity: A linear relationship with the data set size in t-sne embeddings, 2024.

[18] Victoria C Stodden. Reproducible research: Addressing the need for data and code sharing in computational science. 2010.

[19] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

[20] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[21] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

A Example of LLM Prompts

1. in t-SNE, how is the perplexity value calculated?
2. how does the perplexity value balances the local and global structure?
3. Explain how is the Tasic et al. dataset hierarchical in nature?
4. is the learning rate basically the iterations involved in reducing the KL divergence?
5. When visually inspecting an embedding, what is the checklist to go through for a good qualitative analysis?
6. after running experiments, it seems like for a given sample perplexity, if we keep on increasing the full perplexity, the structure of the clusters remains almost stable and similar. What metrics can I use to plot a graph of against some metric to show this.
7. given the subsections below, give an introductory paragraph for this section.
8. which scientific paper first developed the sampling-based approach for t-SNE?
9. how do I compare a grid of embeddings with only one embedding?
10. Why is responsible research important?
11. For my reasearch, what are some key aspects to take care of in terms of responsible research?
12. in the above given method to plot grids, how do I add a title at the top of the embedding plot?
13. give me code to generate a csv file containing cpd values for the sample perplexity and all full perplexities.

B Additional grid embedding

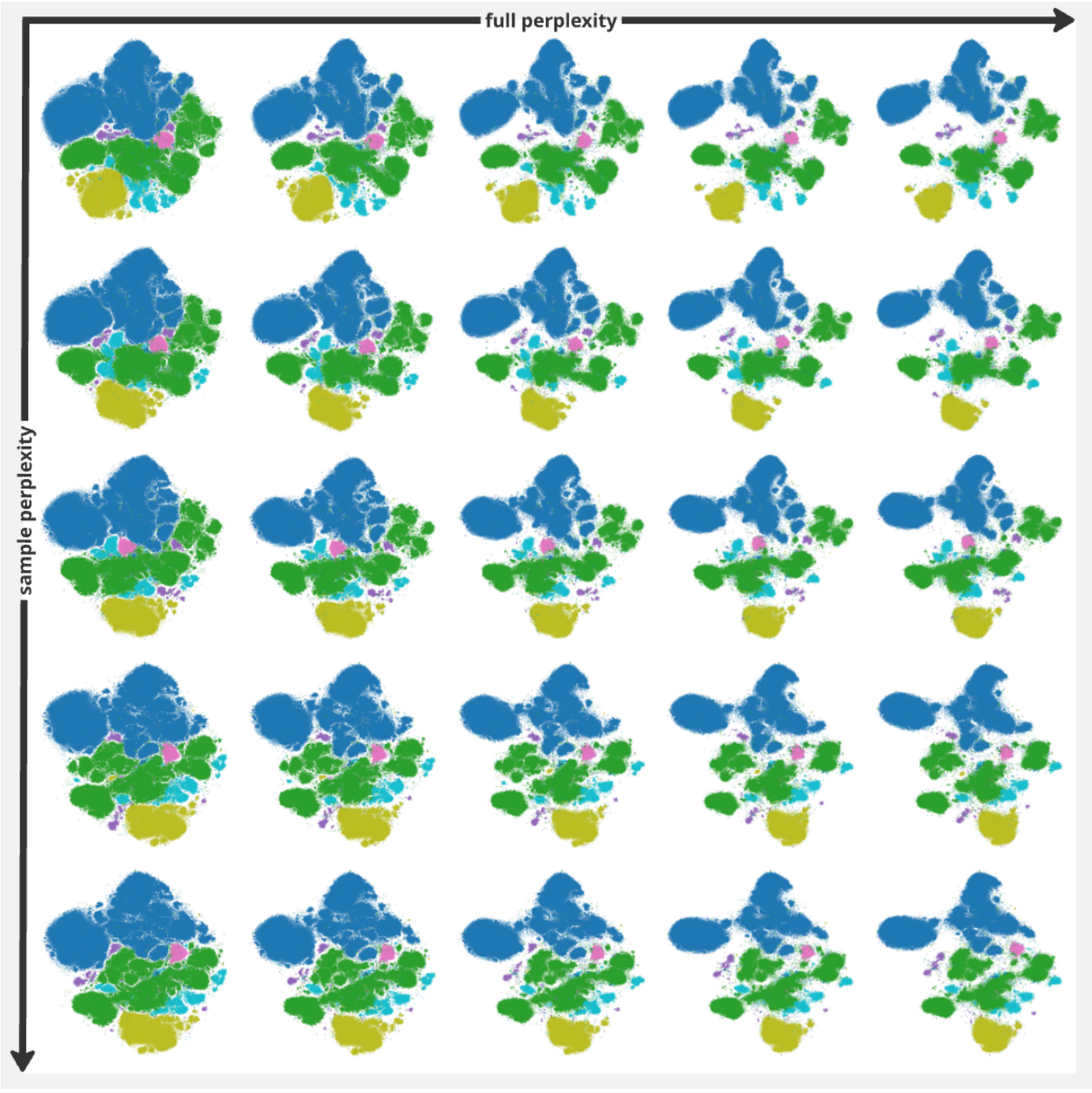


Figure 7: Embedding grid of the Wong dataset created by our approach.