High-Dimensional Data Visualization via Sampling-Based Approaches

Measurement of structural similarity between different embeddings as a way of predicting a suitable perplexity

Radu-Marius Chiriac

Supervisor(s): Klaus Hildebrandt, Martin Skrodzki

[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: Radu-Marius Chiriac
Final project course: CSE3000 Research Project
Thesis committee: Klaus Hildebrandt, Martin Skrodzki, Christoph Lofi

Abstract

Dimensionality reduction techniques, such as t-SNE, are widely used to visualize high-dimensional data and have a crucial role in practical tasks such as biological data exploration [7], anomaly detection [4], or clustering large datasets. However, they are highly dependent on hyperparameters or sampling strategies. This paper investigates whether the structural similarity between sampled and full embeddings can be measured using Procrustes analysis by comparing the structural similarity of the embeddings. This work provides a reproducible framework that quantifies the difference between visualizations produced by sampling t-SNE. These insights provide users a medium to create visualizations with t-SNE without exhaustive experimentation (for example, creating all visualizations), making t-SNE more accessible and reliable.

1 Introduction

High-dimensional data is becoming increasingly common in many fields, such as finance, cybersecurity, or even life sciences, where each data point can have thousands of attributes. Understanding such data can be challenging but rewarding, which is exactly why visualization plays a key role in unfolding properties that can be hard to notice otherwise. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a popular algorithm for projecting high-dimensional data into two or three dimensions for visualization. It is frequently used for clustering large datasets and for anomaly detection in real-world applications such as credit card fraud and network intrusion detection [7], as well as for exploring biological data like single-cell RNA sequencing [4].

High-dimensional data is difficult to interpret and visualize due to its complexity and limited human perception. Dimensionality reduction techniques, such as t-SNE, have become increasingly popular tools for producing qualitative interpretations of 2D or 3D embeddings of such data. However, t-SNE remains a computationally expensive algorithm, especially for large datasets.

Previous work has highlighted the sensitivity of t-SNE to hyperparameters such as perplexity and the choice of initialization [10]. Skrodzki et al. [8] proposed a sampling-based approach, revealing a linear relationship between the perplexity hyperparameter and the sampling ratio. Additionally, methods such as FIt-SNE [6] have accelerated t-SNE for large datasets, enabling experimentation with various sampling strategies. While accelerated implementations like FIt-SNE and better initializations via PCA have significantly improved the runtime of t-SNE on large datasets, they do not directly address the challenges introduced by scale, particularly when full data cannot be stored or processed due to memory or latency constraints. The sample-based t-SNE offers a different approach, instead of generating embeddings from the full data, you use a subset of the dataset. Despite its practical relevance, the structural reliability of embeddings generated from sampled data remains under-explored in the literature.

This paper includes a quantitative way of perplexity selection without needing to exhaustively explore all configuration pairs.

The primary research questions we address are:

- How does the structure of sample-based t-SNE embeddings change across different sampling proportions and perplexity ratios?
- Can we quantitatively measure the similarity between such embeddings using alignment techniques like Procrustes analysis?

Our contributions are as follows:

- We propose a systematic setup for comparing t-SNE embeddings across three different datasets, varying both the sampling proportion and perplexity, inspired by the work of Skrodzki et al. and their findings.
- We introduce a grid-based comparative framework using Procrustes analysis to assess structural similarity across embeddings.
- We provide both qualitative (side-by-side plots) and quantitative (Procrustes disparity values) assessments of embedding stability.
- We briefly discuss an alternative distance metric, namely Wasserstein distance, and its potential advantages in future work.

2 Background

2.1 t-SNE dimensionality reduction

To understand the methods section and the root of the t-SNE sampling issue, first, it is necessary to cover important concepts that lay the groundwork for understanding the approach. T-SNE is an algorithm that converts high-dimensional data points to low-dimensional embeddings. It does this by converting pairwise distances in the original space into conditional probabilities, where the probability $p_{j|i}$ reflects how likely point x_j is to be a neighbor of point x_i.

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

Moreover, t-SNE uses a hyperparameter called *perplexity*, which controls the bandwidths σ_i of the Gaussian kernels used to compute high-dimensional similarities. While perplexity does not directly determine the number of neighbors, it provides a smooth measure of the effective neighborhood size, which refers to the number of data points that significantly influence the position of a given point in the low-dimensional embedding. In efficient implementations of t-SNE, perplexity is also used to sparsify the similarity matrix P, limiting computations to the most relevant neighbors for each point. This parameter ultimately decides how the clusters are formed, for example, lower perplexities will emphasize on the local structure, while a higher perplexity will allow creating broader clusters.

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

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i} \tag{3}$$

Here, $H(P_i)$ is the Shannon entropy of the conditional distribution over neighbors of point x_i. The bandwidth σ_i is chosen via binary search so that the resulting perplexity matches the user-defined value.

Moreover, to construct a low-dimensional embedding, t-SNE minimizes the Kullback–Leibler (KL) divergence between the pairwise similarity distribution of points in the high-dimensional space and that in the low-dimensional embedding. This cost function penalizes cases where similar points in high-dimensional space are placed far apart in the embedding (for example, when the low-dimensional similarity q_{ij} is much smaller than the original p_{ij}). Conversely, placing points which are not similar too close together (when $q_{ij} > p_{ij}$) incurs a smaller penalty [9].

$$\text{KL}(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad \text{where} \tag{4}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{5}$$

While t-SNE is a powerful tool for visualization, it is computationally expensive, which leads us to using sampling-based techniques to reduce runtime and memory requirements by embedding only a subset of the data. This is a result of the iterative optimization process, which leads to a runtime complexity of $\mathcal{O}(n^2)$. and typically requires several hundred steps of gradient descent, further adding to its computational burden.

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ij} - q_{ij}) \cdot \frac{1}{(1 + \|y_i - y_j\|^2)} \cdot (y_i - y_j) \tag{6}$$

Where C is the cost function (KL divergence) and y_i is the position of the i-th point in the low-dimensional space [9]. Sampling offers a possible workaround for the high computational cost of t-SNE by reducing the number of points involved in the process. However, this introduces challenges: the sampled data may not capture the full structure of the dataset, and it is unclear whether a perplexity chosen on the sample generalizes well to the full set [8].

2.2 Procrustes Analysis

Another important aspect that has to be covered is the Procrustes analysis, which is an important tool for the content of this paper as it is our quantitative evaluation metric. What Procrustes analysis does is compare two sets of corresponding points by finding the optimal transformation, which includes translation, scaling, and rotation [3, p. 134]. Ultimately, creating the Procrustes disparity, which measures the residual difference as an error value, if the error is low, then the structure of the embeddings is similar.

Let x_i and y_i denote the 2D coordinates of the i-th data point in two aligned t-SNE embeddings being compared (a reference and a target embedding). We define the Procrustes disparity as:

$$D = \sum_{i=1}^{n} \|x_i - y_i\|^2 = \|X - Y\|_F^2 \tag{7}$$

Here, D is the sum of squared differences between the transformed matrices after optimal scaling, rotation, and translation. Specifically, X and Y are the aligned matrices containing the 2D coordinates of the same subset of n data points, sampled from two different t-SNE embeddings. X corresponds to the reference embedding, while Y corresponds to the embedding being aligned. The Procrustes analysis computes the optimal translation, rotation, and scaling to best align Y to X. The Frobenius norm $\|X - Y\|_F^2$ then quantifies the residual sum of squared differences between the aligned embeddings, yielding the final disparity value, and $\|\cdot\|_F$ denotes the Frobenius norm, which computes the sum of squared differences over all coordinates.

This tool is incredibly powerful for t-SNE because in t-SNE, the absolute positions in the embedding space do not reflect the properties of the points; therefore, transformations such as orientation or scale should not affect the membership of points These transformations do not convey any intrinsic properties of the individual data points, meaning they do not affect the original data values, such as class labels, input features or neighborhood structure, these remain unchanged by geometric transformations such as rotation, scaling or translation. In short, they do affect the global geometry of the embedding, but not the local relationships of t-SNE embeddings.

3 Related Work

The visualization of high-dimensional data through t-SNE [9] has been widely adopted due to its ability to capture local structures. However, the behavior of t-SNE is highly sensitive to its hyperparameters, most notably, perplexity [10]. This has motivated research into better understanding and tuning perplexity values to improve embedding quality and reliability.

A key challenge with t-SNE is that there is no principled way to select perplexity for a given dataset. Belkina et al. [2] proposed the opt-SNE method, which automatically selects perplexity and other parameters using optimization criteria that reflect embedding quality.

Heuristics Used in opt-SNE [2]

- **Early exaggeration stop (EE)** is determined dynamically using the iteration at which the relative Kullback-Leibler divergence change (KLDRC) reaches its local maximum:

$$\mathrm{KLDRC}_N = 100\% \times \frac{\mathrm{KLD}_{N-1} - \mathrm{KLD}_N}{\mathrm{KLD}_{N-1}} \tag{8}$$

  EE is stopped at the next iteration after max(KLDRC).

- **Gradient descent learning rate** η is initialized based on dataset size n and early exaggeration factor α:

$$\eta = \frac{n}{\alpha} \tag{9}$$

- **t-SNE termination** is triggered when the relative improvement in KLD per iteration falls below a threshold:

$$(KLD_{N-1} - \mathrm{KLD}_N) < \frac{\mathrm{KLD}_N}{5000} \tag{10}$$

Their work shows that poor perplexity choices can lead to misleading or distorted embeddings. However, their method assumes access to the full dataset and does not investigate any sampling strategies.

Skrodzki et al. [8] address this issue directly by proposing a linear relationship between dataset size and optimal perplexity, grounded in both empirical evaluation and theoretical intuition. They argue that smaller samples require proportionally smaller perplexity values to preserve structural integrity. While their insights inform the perplexity-sampling trade-off, they do not evaluate how this relationship plays out in embedding stability or pairwise embedding similarity.

To tackle the scalability of t-SNE, Linderman et al. [6] introduced FIt-SNE, a fast interpolation-based approximation of the t-SNE algorithm. FIt-SNE enables the generation of embeddings on large datasets and has become the de facto tool for modern large-scale t-SNE visualizations. Its efficiency makes it suitable for systematic experimentation across various parameter settings. However, while FIt-SNE accelerates full-dataset embeddings, it does not eliminate the computational burden entirely, especially when comparing many configurations, working with extremely large datasets, or operating under memory or latency constraints.

Our work extends these efforts by combining sampling-based perplexity selection with embedding comparison using Procrustes analysis. We provide a systematic framework for evaluating embedding similarity across parameter configurations, helping to clarify how robust t-SNE is to changes in sample size and perplexity.

4 Methods

4.1 Dataset and Preprocessing

We use the MNIST dataset, consisting of 70,000 images of handwritten digits from 0 to 9. Every picture is a vector with 784 dimensions. Additionally, the data is normalised to fall within the [0, 1] range before dimensionality reduction. Moreover, we also use the C. elegans gene expression dataset, which consists of 8,970 cells and 39 gene expression features. Ultimately, the FMNIST dataset was used to further reinforce our findings, a dataset similar to MNIST, but harder to classify.

4.2 Dimensionality Reduction via Sample-Based t-SNE

To investigate the effect of subsampling and perplexity scaling on t-SNE embeddings, we apply a sample-based version of t-SNE where a subset of the full dataset is selected for embedding, and a modified perplexity is chosen proportional to the subset size. Our method is closely aligned with the experiment conducted by Skrodzki et al. [8]. We define the chosen sampling proportions and perplexity ratios for each dataset as follows:

- **Sample proportions (all datasets):** {0.1, 0.4, 0.7, 1.0}
- **Perplexity ratios (MNIST and FMNIST):** {0.00014, 0.00100, 0.00300, 0.02057}
- **Perplexity ratios (C. elegans):** {0.00033, 0.00134, 0.00379, 0.01171}

These specific values were selected to balance quality and interpretability while keeping the total number of embeddings manageable. The sampling proportions range from downsampling (10%) to full dataset use (100%), allowing us to study the effects of sample size across a wide spectrum. The perplexity ratios were derived from values shown to work well in prior work [8], scaled according to the dataset size.

This results in $4 \times 4 = 16$ embedding configurations. While a higher axis (5×5, 6×6) could offer more ground to the analysis, it would exponentially increase the number of comparisons and visualization, complicating both the analysis (i.e, computational costs) and the presentation of our findings (i.e, observing trends). Each configuration produces a 2D embedding using standard t-SNE (FIt-SNE), initialized with PCA [5].

We use this method to better understand overall trends, rows represent increasing sampling proportions from 0.1 to 1.0, while columns represent increasing perplexity ratios. Figure 1 shows a clear pattern: embeddings with low sampling and low perplexity (top-left) tend to visualize poor structure, often forming dense, uninformative clusters. In contrast, embeddings with higher sampling and moderate perplexity (middle and bottom rows, center columns) show well-separated and stable clusters. This suggests a sweet spot where perplexity scales reasonably with the available data, yielding embeddings that capture both local and global structures effectively.

4.3 Embedding Comparison Setup

We compare the 16 embeddings to one another across:

- **Rows:** fixed sample proportion, varying perplexity ratio
- **Columns:** fixed perplexity ratio, varying sample proportion

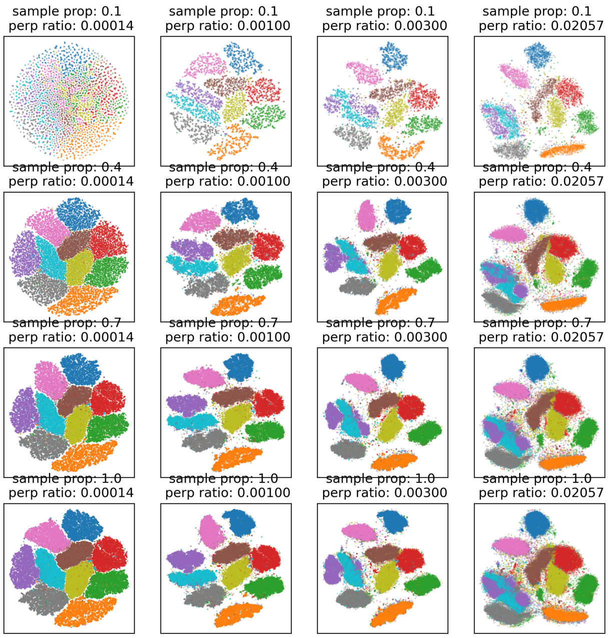


Figure 1: Grid of t-SNE embeddings across different sampling proportions (rows) and perplexity ratios (columns) from the MNIST dataset.

In total, we perform 48 comparisons across our 4 × 4 embedding grid: 24 row-wise (fixed sample, varying perplexity) and 24 column-wise (fixed perplexity, varying sample size). This structure allows us to isolate results and analyze how changes in one parameter affect embedding structure while holding the other constant. Each pair is compared by aligning only the points sampled in common, enabled by our nested sampling strategy in which smaller samples are subsets of larger ones. To compute the Procrustes disparity between two t-SNE embeddings, we first extract the set of data point indices that are shared between the two samples. We then extract the corresponding 2D coordinates resulting in two matrices of equal size representing the common points. These are aligned using Procrustes analysis, which computes the optimal translation, rotation, and scaling. The resulting Procrustes disparity, a single value, helps us quantify the structural difference between the aligned embeddings.

5 Experiments and Results

5.1 Comparison Strategy

Comparisons were grouped as:

- **Row-wise comparisons**: fixed sample proportion, varying perplexity ratio.
- **Column-wise comparisons**: fixed perplexity ratio, varying sample proportion.

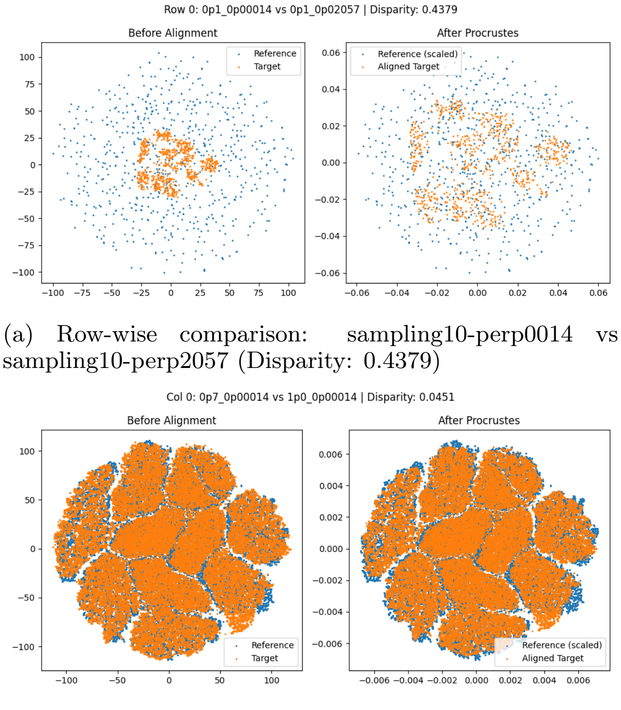
Each pair was aligned based on the intersection of sampled indices to ensure a fair comparison.

5.2 Evaluation

We use the Procrustes disparity as our quantitative evaluation metric. For each comparison, we store the disparity value, number of shared points, and associated plots. Moreover, throughout this section, we will discuss the quantitative and qualitative evaluation of our results, finishing with a brief discussion about Wasserstein distance as an alternative metric.

Qualitative Evaluation

To illustrate how sampling proportion and perplexity interact, we walk through a representative row-wise comparison from our experimental grid. Consider the embeddings generated for 10% of the dataset using two different perplexity ratios: 0.00014 and 0.02057 (i.e., configurations `sampling100-perp0014` and `sampling100-perp2057`).



(a) Row-wise comparison: sampling10-perp0014 vs sampling10-perp2057 (Disparity: 0.4379)

(b) Column-wise comparison: sampling70-perp0014 vs sampling100-perp0014 (Disparity: 0.0451)

Figure 2: Two example comparisons showing structural similarity across parameter configurations drawn from the embedding grid in Figure 1. Top: a high-disparity row-wise case (varying perplexity). Bottom: a low-disparity column-wise case (varying sample size).

These embeddings are visualized in Figure 2a. Before alignment, the higher-perplexity embedding, sampling10-perp2057, appears more globally organized, while the lower-perplexity embedding, sampling10-perp0014, shows tighter but more uninformative close clusters. After Procrustes alignment, some geometric similarity emerges, but local structural differences persist. The resulting Procrustes disparity of 0.4379 quantifies this mismatch, highlighting how sensitive t-SNE can be to perplexity even when sampling is held constant.

In contrast, Figure 2b shows a column-wise comparison between embeddings sampling70-perp0014 and sampling100-perp0014, which differ in sampling proportion but use the same perplexity ratio. These embeddings align remarkably well after transformation, yielding a much smaller disparity of 0.0451, suggesting that with low perplexity, t-SNE embeddings are relatively stable across sample sizes when the structure is sufficiently preserved.

Quantitative Evaluation

We present our findings from conducting several experiments. For each configuration of sampling proportion and perplexity, we ran the embedding three times using different random seeds, 42, 100, and 12. This allows us to account for variability introduced by sampling and provides a more robust comparison across configurations. In each case, the sampled points are different, but the sample size and perplexity remain the same. By averaging the resulting Procrustes disparities, we reduce the impact of outliers and gain a better understanding of the general structural behavior for each setting.

Firstly, we will discuss what we learned after applying the Procrustes analysis metric. Using this metric, we were able to quantitatively evaluate the difference between two different embeddings, and find out how similar by looking at a number instead of relying on visual comparison.

A high disparity would imply there's a bigger difference in how the clusters are formed compared to a smaller disparity. As can be seen in Figures 2a and 2b, we present two comparisons of different embeddings. One is column-wise, while the other is row-wise.

Figure 2 illustrates two such comparisons. The left subplot shows a row-wise comparison where only the perplexity varies, with a relatively high disparity of 0.4379. This reflects a considerable difference in how clusters are formed, especially in the outer regions of the embedding. On the right, we see a column-wise comparison in which the sampling proportion varies but the perplexity is held constant. Here, the disparity is much lower (0.0451), indicating that increasing the sample size does not significantly change the global structure of the embedding, suggesting robustness to sampling when perplexity is scaled appropriately.

Furthermore, after conducting the experiments, we analyzed the structural differences between t-SNE embeddings using Procrustes disparity more generally. The average disparities across multiple runs on the MNIST dataset are presented in Tables 1 and 2, which correspond to column-wise and row-wise comparisons, respectively. While our main analysis focuses on MNIST, further supporting evidence from FMNIST and C. elegans confirms the observed trends (i.e, column-wise vs row-wise analysis). These additional results are included in Appendix B. From these results, we observe that embeddings tend to be more stable when perplexity is held constant and sampling proportions vary, compared to the reverse scenario. This can be seen by observing the heatmap, where a darker color hints at a higher disparity. In general, embeddings as the sampling increases show stability when the perplexity stays the same, while as perplexity increases, the disparity shows greater difference between embeddings of the same sampling size. These findings are especially insightful since they provide a confirmation of how embeddings behave structurally at different sampling and perplexity levels. Moreover, they offer a principled way to guide perplexity selection without exhaustively evaluating all configuration pairs.

One notable finding is that embeddings generated at the lowest sampling proportion (10%) consistently show the highest Procrustes disparity when compared to embeddings at any other sampling level. This can be seen by analyzing Table 1, which shows an obvious increase in disparity when comparisons were made against the 10% sampling column-wise.

This suggests that a sampling rate this low leads to structurally different embeddings, even when using appropriately scaled perplexity values. We hypothesize that this effect is due to the dataset being severely undersampled, which limits the representativeness of the global structure and introduces sparsity that affects neighborhood probability estimation. By repeating the experiment with different random seeds (12 and 100) and datasets (C. elegans and FMNIST), we confirmed that our observations were not specific to one sampling instance, reinforcing the robustness of our findings.

While most comparisons show consistent Procrustes disparities across seeds, the maximum standard deviation observed for MNIST was 0.0761. This happened under conditions of low sampling and low perplexity, where greater embedding variability is expected. In contrast, most stable configurations showed standard deviations below 0.02. On average, the standard deviation across all comparisons was 0.015, indicating a high level of consistency in the embedding generation process.

Alternative Metric: Wasserstein Distance

To complement our analysis of embedding similarity using Procrustes disparity, we computed the Wasserstein distance (Earth Mover's Distance) between pairs of 2D embeddings on MNIST using seed 42. Unlike Procrustes, which relies on pointwise alignment and rigid transformations, Wasserstein distance quantifies the minimal cost of transforming one point distribution into another. This makes it especially suitable for capturing global structural differences, such as spread or density shifts,

Sampling	**Perp. 0.00014**			**Perp. 0.00100**			**Perp. 0.00300**			**Perp. 0.02057**		
Level	40%	70%	100%	40%	70%	100%	40%	70%	100%	40%	70%	100%
10%	0.4463	0.4675	0.4706	0.1214	0.1166	0.1281	0.1280	0.0922	0.1028	0.0680	0.0679	0.0629
40%	–	0.0680	0.0604	–	0.0646	0.0775	–	0.0456	0.0428	–	0.0159	0.0134
70%	–	–	0.0287	–	–	0.0279	–	–	0.0222	–	–	0.0086

Table 1: Column-wise average Procrustes disparities between sampling levels at various perplexity ratios (MNIST, averaged across seeds). Cell color intensity reflects structural dissimilarity.

Perplexity	**Sampling 10%**			**Sampling 40%**			**Sampling 70%**			**Sampling 100%**		
Ratio	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057
0.00014	0.4768	0.4742	0.4456	0.1445	0.1328	0.1713	0.1211	0.0969	0.1589	0.1110	0.0835	0.1612
0.00100	–	0.1690	0.2039	–	0.1203	0.1620	–	0.0882	0.1548	–	0.0765	0.1547
0.00300	–	–	0.1482	–	–	0.1059	–	–	0.0888	–	–	0.1027

Table 2: Similar to Table 1, this shows row-wise average Procrustes disparities between perplexity settings at fixed sampling levels (MNIST, averaged across seeds). Cell color intensity reflects structural dissimilarity.

that are not necessarily aligned geometrically.

The idea of using Wasserstein distance to evaluate embedding quality was inspired by the recent work of Bachmann et al. [1], who introduced Wasserstein t-SNE as a generative approach to constructing embeddings based on optimal transport. While their focus was on embedding construction, we believe it may also be valuable for comparing existing embeddings.

The results, summarized in Table 7 and 8 from Appendix B, show trends consistent with the Procrustes-based analysis: structural dissimilarity increases with lower sampling proportions and lower perplexity values. Furthermore, it supports our finding that structural similarity is better preserved along columns, compared to rows. While the absolute magnitudes differ, the relative relationships between configurations are preserved. This suggests that Wasserstein distance may be a useful alternative tool in assessing embedding stability.

6 Responsible Research

This research focuses on the visualization and quantitative comparison of embeddings created by t-SNE using publicly available data (MNIST, C. elegans), openTSNE, and the sampling t-SNE. As such, it does not involve human subjects, personal data, or other ethically sensitive content.

All experiments are fully reproducible. The full experimental pipeline, including embedding generation, sampling proportions, and Procrustes-based comparisons, was applied independently to MNIST, C. elegans, and FMNIST. We use fixed random seeds (42, 100, 12) during sampling and rely on deterministic variants of t-SNE (FIt-SNE, from openTSNE) to ensure consistent results. The code used for aligning embeddings with Procrustes analysis and visualizing the outcomes is modular and publicly shareable. Our comparisons are based on well-defined procedures, including intersecting sampled indices and using standard mathematical metrics. However, the sampling t-SNE [8], which was used to generate embeddings, is not publicly available.

By making both the source code and processed outputs (CSV files and plots) available, we support transparency, reproducibility, and further extension by other researchers. The methodology can be applied to different datasets or dimensionality reduction techniques without ethical restrictions.

This report made use of an LLM to support the writing process and plotting. The tool was used as follows:

- We consulted the model for feedback with phrasing, LaTeX formatting, figure captions, and flow of written sections.
- Used to debug and refactor plotting code (e.g. matplotlib formatting)
- All prompts used to modify report content are listed in the appendix.
- No AI-generated text has been added without being modified and reviewed.

The prompts used can be found in Appendix A.

7 Conclusions and Future Work

In this work, we investigated how the measurement of structural similarity between different embeddings can be used as a way of predicting a suitable perplexity. Specifically, we focused on a sample-based version of t-SNE, where subsets of the MNIST dataset, C. elegans, and FMNIST were embedded using perplexity values scaled relative to the subset size. Our central research questions were: (1) how does the embedding structure change with different sampling and perplexity configurations, and (2) to what extent can these changes be quantified using Procrustes analysis?

We conducted a 4 × 4 grid of t-SNE embeddings and systematically compared them pairwise across rows (fixed sampling proportion, varying perplexity) and columns (fixed perplexity, varying sampling). By aligning embeddings using Procrustes analysis and reporting the resulting disparity values, we provided a quantifiable measure of structural similarity across 48 comparisons. This revealed that certain regions of the parameter space are more stable than others, particularly at higher sample proportions and moderate perplexities.

One important design choice in our approach is the use of shared sampling indices when comparing embeddings. This ensures that Procrustes alignment operates on consistent point sets across different configurations. While this practice has been used in prior work (e.g., Skrodzki et al. [8]), we explicitly adopt and apply it to create a reproducible framework for grid-based analysis of sample-based t-SNE behavior. This enables a consistent basis for comparison even when the total embedded datasets differ. The visualizations generated also allow for qualitative inspection of structural differences that may not be captured purely by disparity values.

However, several open questions remain. Procrustes analysis relies on Euclidean distance and assumes rigid alignment with scaling, translation, and rotation. This may not fully capture non-linear distortions or differences in distributional structure. To address this, we additionally explored the use of the Wasserstein distance, which compares entire point distributions without requiring one-to-one alignment. Although this metric yielded lower values overall-reflecting its leniency toward global shifts, it still preserved the same qualitative trends observed with Procrustes (i.e, larger differences appeared under low sampling and low perplexity). This suggests that Wasserstein distance could serve as a promising tool for embedding comparison, especially in scenarios where point correspondence cannot be guaranteed. A deeper theoretical and empirical investigation of such distributional metrics remains an important direction for future work.

Future research may also explore:

- Applying this framework to other datasets to potentially find new relationships.
- Further explore the use of Wasserstein distance as a complementary (or alternative) metric for comparing embeddings, including its theoretical properties and practical advantages.

In summary, this work contributes a reproducible pipeline to study how sampling and perplexity affect t-SNE embeddings and offers both visual and numerical tools to compare their outcomes. Moreover, it also provides a way of predicting a suitable perplexity without the need for exhaustive evaluation of all configuration pairs. This paper lays a foundation for more principled evaluation of dimensionality reduction techniques under resource-constrained scenarios.

References

[1] Felix Bachmann, Philipp Hennig, and Dmitry Kobak. Wasserstein t-sne. In M.R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, and G. Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022*, volume 13713 of *Lecture Notes in Computer Science*. Springer, Cham, 2023.

[2] Anna C Belkina, Christina O Ciccolella, Rachael Anno, Rachel 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] Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis*. John Wiley & Sons, 1998.

[4] Infermatic AI. Real-world applications of t-sne in anomaly detection and their use cases. `https://infermatic.ai/ask/?question=What%20are%20some%20real-world%20applications%20of%20t-SNE%20in%20anomaly%20detection,%20and%20what%20are%20their%20respective%20use%20cases?`

[5] Dmitry Kobak and George Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature Biotechnology*, 39:1–2, 02 2021.

[6] George C. Linderman, Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolationâbased tâsne for improved visualization of singleâcell rnaâseq data. *Nature Methods*, 16(3):243–245, 2019.

[7] Scientific Discoveries. What is t-sne plot? `https://www.scdiscoveries.com/blog/knowledge/what-is-t-sne-plot/`.

[8] Martin Skrodzki et al. Navigating perplexity: A linear relationship with the data set size in t-sne embeddings. `https://arxiv.org/abs/2308.15513`, 2023. arXiv preprint arXiv:2308.15513.

[9] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

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

A Appendix: LLM Prompts

This research project utilized large language models (LLMs) to support the writing and formatting processes. The tool was used to improve clarity, automate repetitive scripting tasks, and ensure consistency in presentation. Specifically, the list of used prompts includes:

- "Help me phrase this Procrustes analysis result as a LaTeX formula."
- "Plot Procrustes disparity averages from a CSV using matplotlib."
- "Make this paragraph more concise and academic in tone."
- "Generate heatmap tables in LaTeX from column-wise and row-wise comparisons."
- "How do I format large tables in Overleaf without breaking the layout?"

To ensure no false information was given, all results were analyzed and if needed, computed manually (i.e, values from the table).

B Appendix: Supporting Tables

C. elegans

Column-wise Procrustes Disparities

Sampling Level	**Perp. 0.00033**			**Perp. 0.00134**			**Perp. 0.00379**			**Perp. 0.01171**		
	40%	70%	100%	40%	70%	100%	40%	70%	100%	40%	70%	100%
10%	0.1997	0.2197	0.2266	0.2978	0.2354	0.2363	0.1214	0.1177	0.1219	0.0440	0.0287	0.0296
40%	–	0.0451	0.0426	–	0.1793	0.1748	–	0.0229	0.0198	–	0.0228	0.0230
70%	–	–	0.0220	–	–	0.0365	–	–	0.0118	–	–	0.0058

Table 3: Column-wise average Procrustes disparities at fixed perplexity levels (C. elegans, averaged across seeds). Cell color intensity reflects structural dissimilarity.

Row-wise Procrustes Disparities

Perplexity Level	**Sampling 10%**			**Sampling 40%**			**Sampling 70%**			**Sampling 100%**		
	0.00134	0.00379	0.01171	0.00134	0.00379	0.01171	0.00134	0.00379	0.01171	0.00134	0.00379	0.01171
0.00033	0.3799	0.6514	0.6957	0.2053	0.6630	0.7944	0.1288	0.6651	0.7896	0.0778	0.6725	0.7975
0.00134	–	0.4732	0.5558	–	0.4890	0.6144	–	0.5289	0.6509	–	0.5986	0.7118
0.00379	–	–	0.1293	–	–	0.2367	–	–	0.2072	–	–	0.2159

Table 4: Row-wise average Procrustes disparities at fixed sampling levels (C. elegans, averaged across seeds). Cell color intensity reflects structural dissimilarity.

FMNIST

Column-wise Procrustes Disparities

Sampling Level	**Perp. 0.00014**			**Perp. 0.00100**			**Perp. 0.00300**			**Perp. 0.02057**		
	40%	70%	100%	40%	70%	100%	40%	70%	100%	40%	70%	100%
10%	0.1926	0.2387	0.2442	0.0891	0.0844	0.0869	0.0875	0.0617	0.0602	0.0305	0.0291	0.0281
40%	–	0.0375	0.0356	–	0.0391	0.0453	–	0.0295	0.0289	–	0.0220	0.0211
70%	–	–	0.0190	–	–	0.0206	–	–	0.0146	–	–	0.0115

Table 5: Column-wise average Procrustes disparities at fixed perplexity levels (FMNIST, averaged across seeds).

Row-wise Procrustes Disparities

Perplexity Level	Sampling 10%			Sampling 40%			Sampling 70%			Sampling 100%		
	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057
0.00014	0.3526	0.3003	0.2827	0.0542	0.0837	0.1293	0.0362	0.0725	0.1246	0.0276	0.0725	0.1209
0.00100	–	0.0718	0.1222	–	0.0404	0.0911	–	0.0333	0.0867	–	0.0313	0.0816
0.00300	–	–	0.1482	–	–	0.1304	–	–	0.1183	–	–	0.1097

Table 6: Row-wise average Procrustes disparities at fixed sampling levels (FMNIST, averaged across seeds).

Wasserstein Distance Comparisons

To look beyond Procrustes analysis, we computed Wasserstein distances between embedding configurations. These results are included below for reference. Cell color reflects the intensity of structural shift.

Column-wise Wasserstein

Sampling Level	Perp. 0.00014			Perp. 0.00100			Perp. 0.00300			Perp. 0.02057		
	40%	70%	100%	40%	70%	100%	40%	70%	100%	40%	70%	100%
10%	8.6339	10.7208	12.2054	2.4767	3.2776	3.4097	3.4700	2.4120	2.3228	1.9767	2.0635	2.6605
40%	–	2.1612	3.6408	–	2.3613	2.3336	–	2.8263	2.6825	–	0.7096	1.4056
70%	–	–	1.5723	–	–	2.2411	–	–	0.7221	–	–	0.7903
100%	–	–	–	–	–	–	–	–	–	–	–	–

Table 7: Column-wise Wasserstein distances grouped by perplexity (MNIST, seed 42).

Row-wise Wasserstein

Perplexity Level	Sample 10%			Sample 40%			Sample 70%			Sample 100%		
	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057	0.00100	0.00300	0.02057
0.00014	8.1136	22.8265	46.9828	15.5495	29.4134	56.4428	16.3594	32.6944	59.0281	18.4392	34.5241	61.2934
0.00100	–	14.7950	39.0871	–	14.0475	41.2115	–	16.4173	42.7657	–	16.1323	43.0118
0.00300	–	–	24.5781	–	–	27.3534	–	–	26.5116	–	–	26.9975
0.02057	–	–	–	–	–	–	–	–	–	–	–	–

Table 8: Row-wise Wasserstein distances grouped by sampling level (MNIST, seed 42).