T-SNE EXAGGERATES CLUSTERS, PROVABLY

Anonymous authors
Paper under double-blind review

ABSTRACT

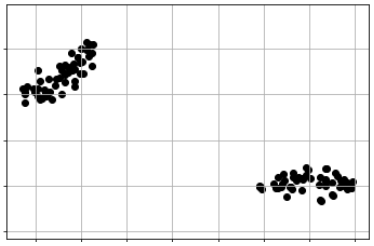
Central to the widespread use of t-distributed stochastic neighbor embedding (t-SNE) is the conviction that it produces visualizations whose structure roughly matches that of the input. To the contrary, we prove that (1) the strength of the input clustering, and (2) the extremity of outlier points, *cannot* be reliably inferred from the t-SNE output. We demonstrate the prevalence of these failure modes in practice as well.

1 INTRODUCTION

t-SNE and related data visualization methods have become staples in modern exploratory data analysis. They just seem to work: practitioners find that these techniques effortlessly tease out interesting cluster structures in datasets. Consequently they are now used ubiquitously in a wide array of fields, ranging from single-cell genomics to language model interpretability (Kobak & Berens, 2019; Petukhova et al., 2025). The practical success of these techniques has naturally piqued some interest in the theoretical community as well.

Existing analysis of t-SNE, for instance, has established that, given high-dimensional data with spherical, well-separated cluster structure, t-SNE outputs a visualization which preserves that cluster structure (Arora et al., 2018; Linderman & Steinerberger, 2019). In other words, t-SNE is provably good at generating *true positives* in its visualization of clusters. Curiously, t-SNE's susceptibility to generate *false positives*, i.e. fabricated clusters in the output visualization, has remained largely unstudied. One should note that this is not a purely academic curiosity, since the interpretation of t-SNE outputs have important consequences downstream in the sciences, like influencing hypothesis generation, experimental design, and deriving scientific conclusions.

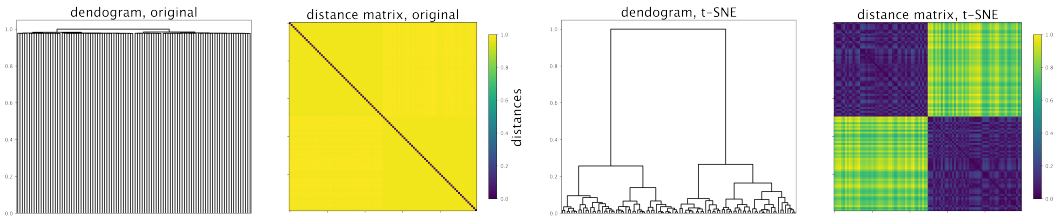
As an illustration of the potential danger of false positives, consider the 2D t-SNE visualization of a 100-dimensional, 100-point dataset (depicted on the right).



Based on this plot, it is tempting to conclude that the input dataset obviously contains two distinct clusters. In this case, one would likely design their subsequent data analysis workflow guided by the two tight, well-separated clusters they see. However a closer examination of the original (high-dimensional) dataset reveals that the situation perhaps may not be as clear-cut. By standard distance-based cluster saliency metrics, the input dataset barely has any gap between the two clusters with respect to the partition that t-SNE so strongly suggests, see Table 1.

Table 1: Clustering scores (with respect to k-means clustering on the t-SNE plot) according to various popular cluster saliency metrics. The range in the second column specifies the possible values that can be attained. A higher value indicates data being highly clustered. Note that $k = 25$ in the last row.

Metric Type	Cluster Score (range)	t-SNE (2D)	Original Data (100D)
Distance-based	Silhouette $[-1, 1]$	.918	.006
	Calinski-Harabasz $[0, \infty]$	5590	1.61
	Dunn Index $[0, \infty]$	3.65	.998
Neighbor-based	k-NN modularity $[0, 1]$	1.0	.95



The interpoint distance matrix along with a hierarchical clustering of the points (via complete-linkage) further elucidates this discrepancy. t-SNE's two-dimensional visualization features a sizable separation between small intra-cluster and large inter-cluster distances. This separation is significantly weaker in the original input data, where interpoint distances are near-uniform. One should not take this to mean that t-SNE is *fabricating* clusters but rather *exaggerating* their separation. Indeed the last row of Table 1 indicates that the t-SNE visualization is consistent with the nearest-neighbor structure of the original data.

Our work formalizes specific limitations of t-SNE in terms of faithfully depicting distance-based information in the input. Our theoretical analysis, suffused with experiments, shows that one should take highly-pronounced well-separated clusters depicted by t-SNE with a grain of salt. Our contributions are as follows:

- **Misrepresentation of cluster distances:** We prove that both strongly-separated and arbitrarily weakly-separated clustered datasets can produce the exact same strongly-separated clustered visualization, see Theorem 3 and Corollary 4. Moreover, we prove that even a slight distance perturbation of inputs can have vastly distinct visualizations, see Theorem 5. We identify the property of t-SNE that explains these peculiar behaviors, and use this understanding to design a targeted adversarial attack that disrupts cluster structure in the output, see Figure 3.
- **Misrepresentation of outliers**: We prove that, regardless of input, the resulting t-SNE output is incapable of depicting extreme outliers, in the sense of depicting one point as substantially far away from all the others, see Theorem 8. In practice, on both synthetic and real datasets, we observe a more concerning phenomenon that faraway outliers are often subsumed into the cluster structure of the bulk of points, see Figures 4 and 5.

While there has been some work investigating the shortcomings of t-SNE in various practical settings (see Section 2.2 for a detailed discussion of the relevant literature), to the best of our knowledge this is the first work which theoretically analyzes some of the key limitations of t-SNE.

2 Related Work

Confidence in the data visualizations produced by t-SNE and related methods is a contentious subject in data science (Marx, 2024; de Bodt et al., 2025). Some argue that these methods have merit in terms of preserving cluster structure and therefore aid in exploratory data analysis, while others warn us about the distortions introduced by these methods.

2.1 Performance Guarantees and Analysis of t-SNE

Shaham & Steinerberger (2017) were among the first to provide a guarantee on the visualization produced by optimal SNE embeddings of well-clustered data. Works by Linderman & Steinerberger (2019) and Arora et al. (2018) refined and extended this analysis, showing that t-SNE outputs produced using gradient descent yield well-clustered visualizations so long as the input is sufficiently well-clustered. The latter work established this guarantee in considerable generality, including cases where the input is sampled from a mixture of well-separated log-concave distributions.

Along with these algorithmic performance guarantee results, there is a line of work that seeks to establish a more fundamental understanding of t-SNE as an optimization problem. Cai & Ma (2022), for instance, characterized the distinct phases of gradient-based optimization of t-SNE, and proved an asymptotic equivalence between the early exaggeration phase of t-SNE and spectral clustering.

Auffinger & Fletcher (2023) proved a consistency result for a continuous analogue of t-SNE, viewing the optimization problem as producing a map between distributions rather than just point sets. Jeong & Wu (2024) and Weinkove (2024) studied the gradient flow of t-SNE. The former showed mild assumptions under which optima exist, and the latter showed that, even in cases where the gradient flow diverges the relative interpoint distances stabilize in the limit.

2.2 Weaknesses and Criticisms

Bunte et al. (2012) were among the first to investigate the potential shortcomings of using KL-divergence in a t-SNE visualization and proposed a generalization to other divergences that may be better suited for specific datasets and user needs. Building upon the precision-recall framework of Venna et al. (2010), Im et al. (2018) extended this result and explored specific intrinsic structures within data that may be less suited for t-SNE. They concluded that while t-SNE is more attuned to reveal intrinsic cluster structure, it usually fails to reveal intrinsic manifold structure.

In terms of analyzing cluster structure specifically, Yang et al. (2021) provided empirical evidence that t-SNE visualizations are prone to *false negatives*. They presented a selection of well-clustered real-world datasets which t-SNE embeddings, even with reasonable parameter-tuning, do not seem to represent faithfully. They also showed that these practical datasets do not abide by the theoretical cluster separation conditions that are required by Arora et al. (2018) analysis. Chari & Pachter (2023) argued that t-SNE and UMAP are unreliable tools for exploratory data analysis. Taking single-cell genomic data as an important real-world example, they provided systematic empirical evidence that these embeddings suffer high distortion, and often misrepresent neighborhood and cluster structure. Curiously, to the best of our knowledge, there is no systematic theoretical study investigating false positive behavior of t-SNE.

More recently, Snoeck et al. (2025) provided theoretical evidence that, not just t-SNE, but any embedding technique that attempts to visualize data in constant dimensions is bound to misrepresent neighborhood structure in most datasets.

3 Preliminaries

Given an input dataset[1] $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^D$, the goal of t-SNE is to come up with an embedding $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ (where $d \ll D$, typically $d = 2$) that approximately maintains the neighborhood structure in X. t-SNE accomplishes this by assigning affinities to input data points which encode how likely an input point is to be a neighbor to a given point. The goal then is to find a configuration of the embedded points Y that induces a similar neighborhood affinity. Specifically, let $P = P(X) \in \mathbb{R}_+^{n\times n}$ and $Q = Q(Y) \in \mathbb{R}_+^{n\times n}$ be the input and embedded *affinity matrices* describing the pairwise neighborhood similarities in the input and output, respectively. t-SNE constructs P by first computing neighborhood affinities for each point i defined as (for any $j \neq i$)[2]

$$P_{j|i}(X;\sigma_i) := \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k\neq i}\exp(-\|x_j - x_k\|^2/(2\sigma_i^2))}, \tag{1}$$

where σ_i encodes the (point-dependent) neighborhood scaling. It is worth noting that $P_{\cdot|i}$ is a valid probability distribution over $[n]$. The matrix P is then constructed based on a crucial parameter called the perplexity (denoted as ρ and taking values in $[1, n-1]$), as follows:

(1) For each i, select the neighborhood scale σ_i^* such that the entropy of the neighborhood distribution $P_{\cdot|i}(X;\sigma_i^*)$ is $\log\rho$.

(2) Define $P = [P_{ij}]_{i,j\in[n]}$ where $P_{ij} := \frac{1}{2n}(P_{i|j}(\sigma_j^*) + P_{j|i}(\sigma_i^*))$ if $i \neq j$ and zero otherwise.

To avoid the so-called *crowding problem* (see Van der Maaten & Hinton (2008) for details), the output affinity matrix Q is computed based on a t-distribution. Specifically, for $i \neq j$

$$Q_{ij}(Y) := \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k,l;k\neq l}(1 + \|y_k - y_l\|^2)^{-1}} \qquad Q_{ii} = 0. \tag{2}$$

[1] Without loss of generality, we shall assume that the input dimension $D = n - 1$.

[2] When X and σ_i^* are clear from context, we will often drop it from the notation.

As indicated before, the objective then is to minimize the gap between the input and output affinities P and Q. This is accomplished by minimizing relative entropy (KL-divergence) between the P and Q affinities (viewed as probability distributions).

$$\text{minimize}_Y \ \mathcal{L}_X(Y) := \mathrm{KL}(P(X)\|Q(Y)) = \sum_{\substack{i,j \\ i\neq j}} P_{ij}(X) \log\Big(\frac{P_{ij}(X)}{Q_{ij}(Y)}\Big).$$

This highly non-convex objective is usually optimized by initializing at a good starting point via an *early exaggeration phase*, followed by performing standard gradient descent methods and returning an embedding Y that corresponds to a local minimum of the objective. Our central task is to study the nature of the these (local minimum) embeddings returned by t-SNE and their relation to the space of input datasets.

Definition 1. *For an n-point dataset $X \subset \mathbb{R}^{n-1}$ and perplexity parameter $\rho \in [1, n-1]$, define*

$$\text{t-SNE}_\rho(X) := \{Y \subset \mathbb{R}^d : \nabla_Y \mathcal{L}_X(Y) = 0\}$$

as the set of outputs $Y \subset \mathbb{R}^d$ that are stationary to the t-SNE objective on a given input X.

Furthermore, for a set of n-point datasets $\mathcal{X}_n$, we define: $\text{t-SNE}_\rho(\mathcal{X}_n) = \bigcup_{X\in\mathcal{X}_n} \text{t-SNE}_\rho(X)$. If $\mathcal{X}_n$ is the set of all *n-point datasets, we denote $\text{t-SNE}_\rho(\mathcal{X}_n)$ as $\mathsf{Im}(\text{t-SNE}_{\rho,\mathrm{n}})$ to indicate the entire image of the t-SNE map.*

All the supporting proofs for our formal statements can be found in the Appendix, and the code related to our experimental demonstrations is available on Github at https://github.com/anon594/iclr26_submission8125.

4 Misrepresentation of Cluster Distances

Previous works by Linderman & Steinerberger (2019) and Arora et al. (2018) have identified that clustered inputs induce clustered t-SNE visualizations in the sense that sufficiently well-separated Gaussian-shaped clusters in the input must produce corresponding well-separated clusters in the visualization. A key question for practitioners left unanswered by these analyses is: when does a clustered output imply a clustered input? More generally, what information can be deduced about the input given a visualization? We answer this question by providing theoretical and practical evidence that the strength of cluster separation in the input, unfortunately, cannot be reliably inferred from the low-dimensional visualization.

To quantify strength of cluster separation in a dataset, we employ well-known distance-based cluster indices such as the average silhouette score (Rousseeuw, 1987), the Calinski-Harabasz index (Caliński & Harabasz, 1974), and the Dunn index (Dunn, 1974). For sake of readability, we focus on presenting our results with respect to the average silhouette score. Our results hold identically for the other indices as well (see Appendix A).

Definition 2. *Given a partition $C_1 \sqcup C_2 \sqcup \cdots \sqcup C_k = [n]$ of n points $\{x_1, \ldots, x_n\} = X$, the **silhouette score** of a point x_i (w.r.t. the partition), denoted $\mathcal{S}(i)$, is the normalized difference between the average within- and the closest across-cluster distances from x_i:*

$$\mathcal{S}(i) := \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \qquad a(i) := \sum_{j\in C^{(i)}} \frac{\|x_i - x_j\|}{|C^{(i)}| - 1} \qquad b(i) := \min_{\substack{m\in[k] \\ C_m \neq C^{(i)}}} \sum_{j\in C_m} \frac{\|x_i - x_j\|}{|C_m|},$$

*where $C^{(i)}$ is the cluster to which i belongs. Note that if $|C^{(i)}| = 1$, then $\mathcal{S}(i)$ is defined to be zero. The **average silhouette score** then is simply the average across all points in X:*

$$\bar{\mathcal{S}}(X; C_{m\in[k]}) := \frac{1}{n} \sum_{i\in[n]} \mathcal{S}(i).$$

Observe that the (average) silhouette score ranges from -1 to 1 with higher scores reflecting large separation between clusters relative to cluster diameter. A score of zero reflects minimal separation between clusters, while negative values reflect cluster overlaps.

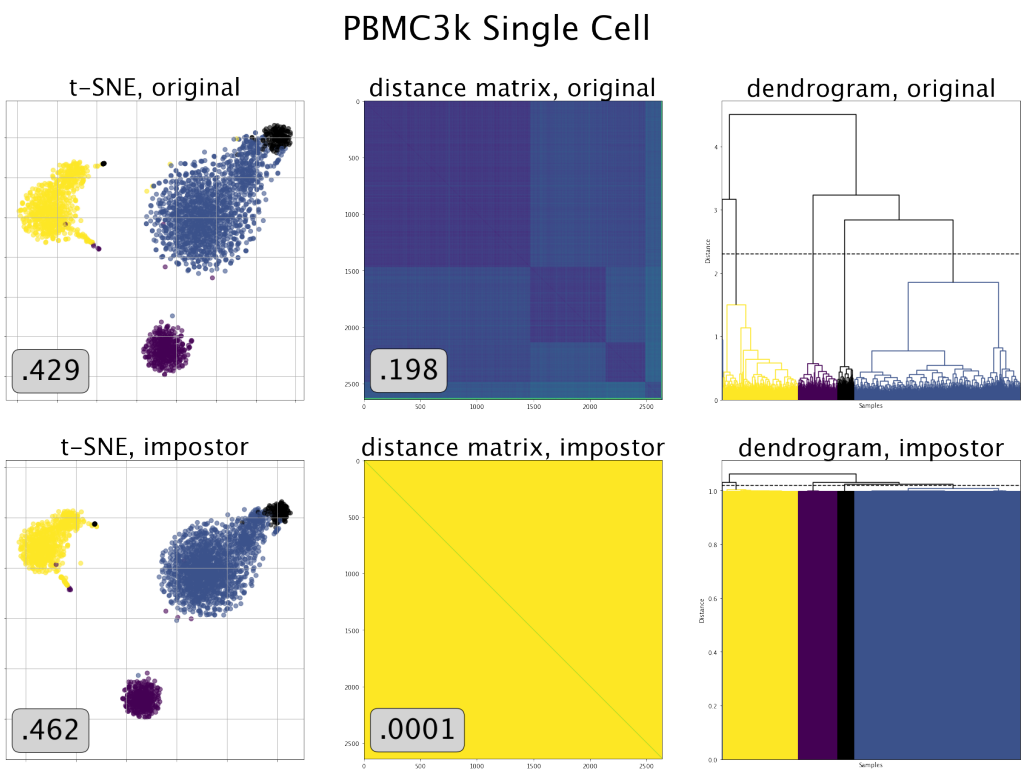


Figure 1: Visualizations of single-cell data (top row) versus an impostor dataset with arbitrarily small cluster separation (bottom row). Based on the 2D t-SNE visualization (left column), it is difficult to distinguish which dataset (real or impostor) may have produced the plot. Plotting the input interpoint distance matrices (middle column) suggests that the clusters in the impostor dataset are significantly less separated than in the original dataset. The corresponding dendrograms (right column, produced using the Ward's method) further elucidate the relative strength of pairwise distances. It is worth emphasizing that the impostor dataset retains the relative rank ordering of the pairwise distances (and therefore the ordering of nearest neighbors), and only distorts the distances to make the differences much finer. Note that the color coding in all of the scatterplots corresponds to a stable cut in the hierarchical clustering given by the dendrogram of the original dataset, and the bottom left labels correspond to silhouette scores with respect to that clustering.

4.1 Different Inputs, Same Output

Defining the strength of a clustering with respect to this cluster index, we show that any stationary t-SNE output (where the clusters may be well-separated or otherwise) can be produced by an input with minimal distance separation between the clusters:

Theorem 3. *Fix any $n > k > 1$, and n-point dataset $X \subset \mathbb{R}^{n-1}$ with partition $C_1 \sqcup \cdots \sqcup C_k = [n]$ such that $|C_{m\in[k]}| > 1$ and $\bar{\mathcal{S}}(X; C_{m\in[k]})$ is well defined. For all $0 < \epsilon \leq 1$, there exists n-point dataset $X_\epsilon \subset \mathbb{R}^{n-1}$ such that*

$$\bar{\mathcal{S}}(X_\epsilon; C_{m\in[k]}) = \epsilon \cdot \bar{\mathcal{S}}(X; C_{m\in[k]}),$$

yet, for any $\rho \in [1, n-1]$:

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X_\epsilon).$$

It is important to understand the implications of this result. For any high-dimensional dataset X (that contains well-separated clusters), we can always find an impostor dataset X_ϵ with minimal distance separation such that *all* t-SNE (local as well as global) stationary points of X and X_ϵ match perfectly! In other words it is *impossible* to distinguish between X and X_ϵ based on the low-dimensional t-SNE visualization.

As a consequence, the same well-separated clustered visualization can be produced by a sequence of impostor datasets containing clusters ranging from well-separated to minimally separated,

Corollary 4. *For all $n \geq 4$ even, and partition $C_1 \sqcup C_2 = [n]$ such that $|C_1| = |C_2| = \frac{n}{2}$. There exist a sequence of n-point datasets in $\mathbb{R}^{n-1}$, $\{X_\epsilon\}_{0<\epsilon\leq 1}$, with*

$$\bar{\mathcal{S}}(X_\epsilon; C_1, C_2) = \epsilon$$

such that for any $\rho \in [1, n-1]$, we have $Y \in \bigcap_{0<\epsilon\leq 1} \text{t-SNE}_\rho(X_\epsilon)$ with

$$\bar{\mathcal{S}}(Y; C_1, C_2) = 1.$$

The above shows that Y, a perfectly clustered visualization according to silhouette score, is a local (and global, see the proof in Appendix) minimizer for any member of a set of inputs of *arbitrary* silhouette score. Thus, even from a visualization which is *perfectly* clustered, the strength of the input's cluster structure cannot be inferred.

Note that the existence of an impostor X_ϵ is not just theoretical; it can be constructed practically as well (see Appendix A.7 for an explicit construction). Hence this phenomenon can be demonstrated in real-world scenarios, see Figure 1. In this case, we select a preprocessed version of the well-known PBMC3k single-cell genomics dataset (2638 points, 50 dimensions; 10x Genomics (2019)) as X. We show that there is an "impostor dataset" X_ϵ that is essentially indistinguishable from the real dataset in terms of its 2D t-SNE visualization, yet has a much weaker cluster separation than the original dataset. The difference between impostor and original can be quantified by silhouette score and visualized in terms of the interpoint distance matrix and dendrogram.

It is worth emphasizing that while the distance information in the original and impostor dataset is dramatically different, the *ordinal* relationship between neighbors is identical between the datasets—as demonstrated by the corresponding structure in the dendrograms. This observation indicates that t-SNE is not fabricating clusters in its visualization of the impostor, but rather exaggerating their salience.

4.2 Diverse Outputs, Similar Inputs

The previous section established that inputs with vastly different metric structure can yield the exact same t-SNE output. We continue with a complementary result: that near-isometric inputs can yield an arbitrarily rich diversity of outputs.

Theorem 5. *Fix any $n \geq 2$ and $\rho \in [1, n-1]$. For all $\epsilon > 0$ and all $Y, Y' \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$, there exists n-point datasets $X = \{x_1, \ldots, x_n\}$ and $X' = \{x'_1, \ldots, x'_n\} \subset \mathbb{R}^{n-1}$ such that $\forall i \neq j$*

$$1 - \epsilon \leq \frac{\|x_i - x_j\|^2}{\|x'_i - x'_j\|^2} \leq 1 + \epsilon,$$

yet $Y \in \text{t-SNE}_\rho(X)$ and $Y' \in \text{t-SNE}_\rho(X')$.

Thus even minor interpoint distance perturbations of the input dataset can develop into massive changes in the visualization. Figure 2 demonstrates this phenomenon quite clearly. We start with a dataset X that is a regular unit simplex (all pairwise distances are unit length). By systematically perturbing the input X ever so slightly ($\epsilon \leq 0.01$), t-SNE produces strikingly different outputs.

The key observation behind our main Theorems 3 and 5 is the simple yet seemingly counter-intuitive fact that t-SNE is not only invariant under multiplicative scaling of the input squared distances, but also *additive* shifts thereof – a property also investigated by Lee & Verleysen (2011; 2014). Specifically given a dataset $X = \{x_1, \ldots, x_n\}$, for any dataset $X' = \{x'_1, \ldots, x'_n\}$ and $C \in \mathbb{R}$ such that, $\|x'_i - x'_j\|^2 = \|x_i - x_j\|^2 + C \geq 0$ for $i \neq j$, we have $\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X')$ (see Lemma 15 for a formal statement, and Lee & Verleysen (2011), Section 4). As a consequence, for any input dataset, we can simply pump up the interpoint distances and construct an impostor dataset which has the same visualization profile but is arbitrarily close to a regular simplex (and hence is arbitrarily unclustered)[3]. This observation also leads to the following seemingly bizarre fact.

[3] See Algorithm 1 for a formalization of this process.

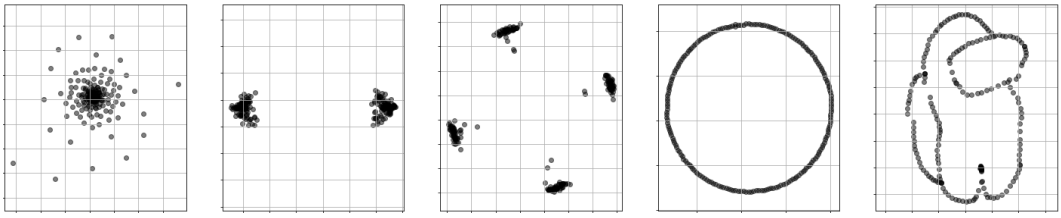


Figure 2: Various different 2D t-SNE visualizations produced by adversarial perturbations of a 200-point unit regular simplex. Each pair of perturbations satisfies the conditions of Theorem 5 for $\epsilon = 0.01$.

Lemma 6. *Fix any $n \geq 2$ and $\rho \in [1, n-1]$. For any $\epsilon > 0$, define the set of ϵ-perturbations of a unit simplex as $\Delta_\epsilon := \{X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1} : \forall i \neq j, \|x_i - x_j\|^2 \in [1-\epsilon, 1+\epsilon]\}$. Then, for all $\epsilon > 0$*

$$\mathsf{Im}(\text{t-SNE}_{\rho,\text{n}}) = \text{t-SNE}_\rho(\Delta_\epsilon).$$

In other words there is a set of datasets Δ_ϵ arbitrarily close to a regular unit simplex that generates *all* possible stationary t-SNE outputs! This result indicates that t-SNE outputs are highly unstable on near-simplex inputs (cf. Figure 2), which has real-world consequences since many high-dimensional datasets fall into this regime (Beyer et al., 1999; Aggarwal et al., 2001) due to the concentration of measure phenomenon (Ledoux, 2001).

4.2.1 A Simple Adversarial Attack: Poison Points

The previous lemma tells us that on intrinsically high-dimensional (near-simplex) data, small perturbations of all the interpoint distances can have outsize effects on the t-SNE output. We observe that such datasets are susceptible to a much simpler adversarial attack: namely, the insertion of just a *single* data point, placed between clusters.

Consider a dataset X sampled from a mixture of two high-dimensional Gaussians. t-SNE, as expected, reveals the two underlying clusters (cf. Figure 3, first panel). However, we can add just a single "poison point" to X and destroy the clustered visualization (see Figure 3 second panel). This failure mode of t-SNE is also observed on a real high-dimensional datasets (see Figure 5 left vs. center).

The success of the poison point attack can be attributed to additive invariance as follows. Given an input dataset in Δ_ϵ from a clustered, high-dimensional distribution, the squared interpoint distances occupy a tight band between $1-\epsilon$ and $1+\epsilon$ due to concentration of measure. Since t-SNE is invariant under additive scaling, the dataset appears identically as if all the (square) distances are in the range $[0, 2\epsilon]$. Thus, from t-SNE's perspective, the variation between inter-cluster distance ($\approx 2\epsilon$) and intra-cluster (≈ 0) is large. However, when the single point is added at the mean, the minimum (non-squared) distance from any point to the rest of the set is approximately halved. As a result, almost all distances remain in the range $[1-\epsilon, 1+\epsilon]$, but, as t-SNE sees it, the effective inter-cluster ($\approx (1+\epsilon) - \frac{1}{4}(1-\epsilon) = \frac{3}{4} + \frac{5}{4}\epsilon$) and intra-cluster ($\approx (1-\epsilon) - \frac{1}{4}(1-\epsilon) = \frac{3}{4} - \frac{3}{4}\epsilon$) gap has been reduced, causing the cluster structure to go unrecognized in some cases.

We observe that single poison point attacks can destroy or significantly weaken cluster structure for very large datasets, see Figure 11. While the attack is not guaranteed to work, we find that its efficacy seems to scale with the number of points added, though care must be taken to ensure that poison points are far enough away from each other so as not to form their own cluster.

We explore this phenomenon further in the next section, where we contrast it with t-SNE's strikingly indifferent response to the injection of outlier points.

5 Misrepresentation of Outliers

Most analysis on t-SNE, including the previous section, is concerned with whether it faithfully depicts global structure, specifically cluster structure. In this section, we consider how t-SNE repre-

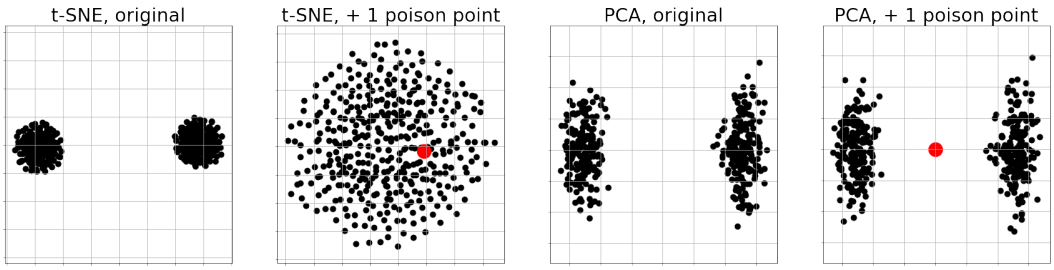


Figure 3: t-SNE versus PCA's radically different responses to the injection of a single "poison" point in the input dataset. The original dataset, visualized in panels 1 and 3, consists of 400 points sampled from a mixture of two Gaussians in $\mathbb{R}^{2000}$. The poison point is then placed at the mean of the previously sampled points; the resulting 401 point dataset is visualized in panels 2 and 4. Note that this behavior persists for larger datasets, see Appendix B.1.

sents points that drastically deviate from the global structure: namely, *outliers*. It is natural to hope that data visualization methods can enable the identification of outliers. Unfortunately, we find that t-SNE may arbitrarily suppress the severity of outliers present in the input dataset.

This phenomenon has been observed empirically in prior work, though to our knowledge we are the first to formalize it (Schubert & Gertz, 2017). An intuitive explanation of t-SNE's response to outliers can be made based on the asymmetry of the input and output affinity matrices of t-SNE. Roughly speaking, the input affinity behaves like a normalized, symmetrized nearest neighbor graph, whereas the output affinity behaves more like a radius neighbors graph. This means the output affinity is optimized to represent the outlier point in close proximity with at least some points, even if it was extremely far from those points in the input.

To begin to formalize this observation, we provide a geometric definition of an outlier.

Definition 7. *Fix $X \subset \mathbb{R}^D$, $x_0 \in X$, and $\alpha \in \mathbb{R}_+$. We say X is an (α, x_0)-**outlier configuration** if there exists a hyperplane separating x_0 and $X \setminus \{x_0\}$ with margin width at least*

$$\alpha \cdot \max\{1, \operatorname{diam}(X \setminus \{x_0\})\},$$

*Define the **outlier number** of a dataset, denoted $\alpha(X)$, as the largest α for which there exists $x_0 \in X$ such that X is an (α, x_0)-outlier configuration.*

This definition can be generalized to accommodate more than one outlier, but for the purposes of theoretical analysis we consider just one. Note that the outlier extremity α is defined relative to the diameter of the rest of the points, unless that diameter is below 1. The choice of a threshold here is important and intuitive: it allows us to have a suitable notion of outlier in extreme cases such as when $\operatorname{diam}(X \setminus \{x_0\}) = 0$.

Our main theorem establishes that any stationary t-SNE output, *regardless of its input*, is incapable of depicting extreme outliers.

Theorem 8. *Fix $n > 2$ and $\rho \in [1, n-1]$. Let $Y = \{y_0, y_1, \ldots, y_{n-1}\} \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$ be a stationary t-SNE embedding. Without loss of generality let y_0 be the outlier point. Then we have:*

$$\alpha(Y) = \alpha(Y, y_0) \leq \sqrt{1 + \left(1 + \frac{2}{n-2}\right)\left(\frac{12}{1 + \sum_{i=1}^{n-1} P_{0|i}(X)}\right)} = 3.602 + o(1)$$

for all $X = \{x_0, x_1, \ldots, x_{n-1}\}$ such that $Y \in \text{t-SNE}_\rho(X)$.

The result is proven via analysis of the t-SNE gradient: we argue that if the outlier is too far away, its gradient is nonzero, thus violating stationarity. Key to this analysis is a comparison between the aggregate behavior of the outlier point's affinities in the input versus the output; in other words, the comparison between $\sum_{i=1}^n P_{i0}$ and $\sum_{i=1}^n Q_{i0}$. This is where the fundamental asymmetry of t-SNE comes in. While the latter is dependent on the position of the outlier point y_0, per Lemma 18, the former has a lower bound of $1/(2n)$ due to the normalization of the conditional affinity probabilities.

The input-agnostic nature of this result is striking: even if the input is an extreme outlier configuration, a t-SNE output cannot depict its extremity past roughly $\alpha = 3.6$. This behavior stands in

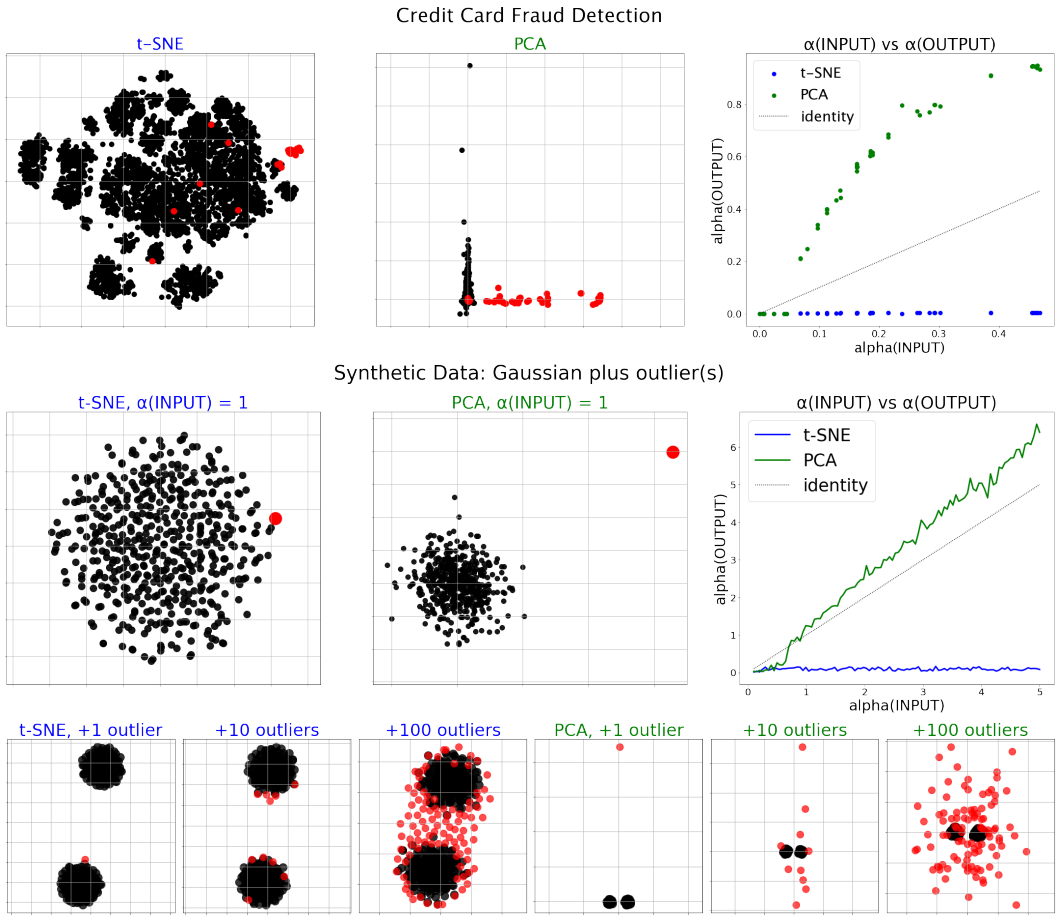


Figure 4: t-SNE's response to α-outliers, compared with PCA. Top row: given data monitoring financial activity ($n = 5050$, $D = 30$) where one percent of users are committing fraud, PCA succeeds and t-SNE fails at representing the fraudulent users as outliers. Note that all of the fraudulent users register as ($\alpha > 0$)-outliers with respect to the regular users; in the top right we show how t-SNE and PCA represent those α-values in their output. Middle row: a similar analysis on a synthetic dataset comprised of a Gaussian sample plus an outlier. Bottom row: mixture of two Gaussians plus 1, 10, and 100 outliers. t-SNE shows the outliers are essentially part of the cluster structure, while for PCA the outliers overtake the structure of the embedding.

stark contrast to that of principal component analysis (PCA), as shown in Figure 4 on both real and synthetic data models. PCA tends to preserve the α outlier number, while t-SNE seldom depicts outliers past $\alpha > 0.2$ in practice, and sometimes even depicts them as within the convex hull of the rest of the points (hence $\alpha = 0$). Furthermore, when faced with multiple outliers, (Figure 4, bottom) t-SNE gracefully accommodates them into the global structure of the bulk of the data.

Our result suggests that t-SNE is the wrong tool to use in situations involving outlier detection. Consider, for instance, a dataset of financial transactions where the goal is to detect fraudulent user, studied by Pozzolo et al. (2015). In this dataset, only 0.172% percent of the points (492 out of $284,807$) are fraudulent and by many standard statistical metrics register as outliers. Comparing the t-SNE and PCA plots on a random representative subset of this data (5050 points, of which 50 are fraudulent), we see that t-SNE mixes the frauds with the bulk of the points while PCA keeps them separated for the most part, see Figure 4, top row.

Finally, note the distinction between t-SNE's muted response to outliers and its dramatic sensitivity to poison points. We illustrate this distinction on a dataset of BBC news articles (Greene & Cunningham, 2006), see Figure 5. Given RoBERTa (Liu et al., 2019) sentence embeddings of these articles ($n = 2225$, $D = 1024$), we find that injecting 220 poison points (see Appendix B.1 for the explicit construction) can halve the silhouette score of the t-SNE embedding with respect to the ground-truth labelling, whereas injecting 1100 large-α-outliers slightly improves the silhouette score.

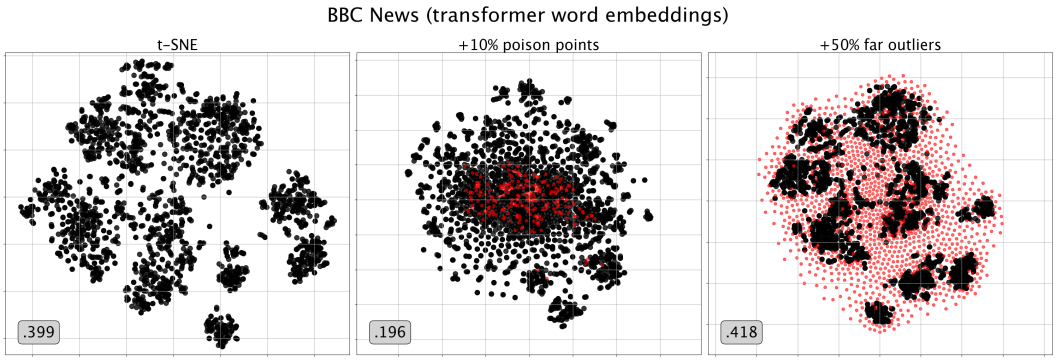


Figure 5: t-SNE's susceptibility to poison points in contrast with its muted response to outliers, on the BBC news article classification dataset. The bottom left label denotes silhouette score of the original points (without the injected points) with respect to the true labels (business, entertainment, politics, sport, tech).

6 Discussion

Our study of t-SNE has established in considerable generality that one cannot infer the *degree* of cluster separation or the extremity of outliers from a t-SNE plot, see Theorems 3, 5, and 8. The proofs and intuitions behind these statements guided us to the surprising empirical observation that one cannot even infer the *existence* of clusters in the presence of outliers. In particular, the injection of a small subset of adversarially chosen points can largely mask the cluster structure, while sizable injections of outlier points are reliably masked within the cluster structure, see Figures 3, 4, 5, and 13.

We have identified two properties of t-SNE that give rise to these idiosyncratic behaviors: (1) additive invariance with respect to the squared interpoint distances, and (2) the asymmetry between the input and output affinity matrices. While we have uncovered significant false positive failure modes that arise from these properties, we cannot completely rule out their utility. For instance, though additive invariance may lead to exaggerated clusters, its robustness vis-a-vis high-dimensional random noise has been discussed in prior work, see e.g. Lee & Verleysen (2011; 2014); Karoui (2010); Karoui & Wu (2015); Landa & Cheng (2023)

t-SNE belongs to a wide selection of data visualization techniques that are yet to be understood fully (McInnes et al., 2018; Jacomy et al., 2014; Tang et al., 2016; Amid & Warmuth, 2019). Our preliminary experiments (see Appendices A.3 and B.2) suggest that the failure modes discussed in this paper may extend to other force-based dimension reduction techniques. Our hope is that this work inspires the reader to explore this fascinating landscape further and pursue the essential question: what can be provably deduced from a visualization?

References

10x Genomics. PBMCs from C57BL/6 mice (v1, 150x150), Single Cell Immune Profiling Dataset by Cell Ranger v3.1.0. https://www.10xgenomics.com/, 2019.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. *International Conference on Database Theory (ICDT)*, pp. 420–434, 2001.

Ehsan Amid and Manfred K Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *Computing Research Repository (CoRR)*, abs/1910.00204, 2019.

Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-SNE algorithm for data visualization. *Conference on Learning Theory (COLT)*, pp. 1455–1462, 2018.

Antonio Auffinger and Daniel Fletcher. Equilibrium distributions for t-distributed stochastic neighbour embedding. *Computing Research Repository (CoRR)*, abs/2304.03727, 2023.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? *International Conference on Database Theory (ICDT)*, pp. 217–235, 1999.

Kerstin Bunte, Sven Haase, Michael Biehl, and Thomas Villmann. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

T Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research (JMLR)*, 23(1):13581–13634, 2022.

T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3 (1):1–27, 1974.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8), 2023.

Cyril de Bodt, Alex Diaz-Papkovich, Michael Bleher, Kerstin Bunte, Corinna Coupette, Sebastian Damrich, Enrique Fita Sanmartin, Fred A. Hamprecht, Emőke Ágnes Horvát, Dhruv Kohli, Smita Krishnaswamy, John A. Lee, Boudewijn P. F. Lelieveldt, Leland McInnes, Ian T. Nabney, Maximilian Noichl, Pavlin G. Poličar, Bastian Rieck, Guy Wolf, Gal Mishne, and Dmitry Kobak. Low-dimensional embeddings of high-dimensional data. *Computing Research Repository (CoRR)*, abs/2508.15929, 2025.

J. C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1): 95–104, 1974.

Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. *International Conference in Machine Learning (ICML)*, 2006.

Daniel Jiwoong Im, Nakul Verma, and Kristin Branson. Stochastic neighbor embedding under f-divergences. *Computing Research Repository (CoRR)*, abs/1811.01247, 2018.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One*, 9(6):e98679, 2014.

Seonghyeon Jeong and Hau-Tieng Wu. Convergence analysis of t-SNE as a gradient flow for point cloud on a manifold. *Computing Research Repository (CoRR)*, abs/2401.17675, 2024.

Noureddine El Karoui. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5), 2010.

Noureddine El Karoui and Hau-Tieng Wu. Connection graph laplacian methods can be made robust to noise. *The Annals of Statistics*, 2015.

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

Boris Landa and Xiuyuan Cheng. Robust inference of manifold density and geometry by doubly stochastic scaling. *SIAM Journal on Mathematics of Data Science*, 5(3):589–614, 2023.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2001.

John A. Lee and Michel Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 2011.

John A. Lee and Michel Verleysen. Two key properties of dimensionality reduction methods. *Computational Intelligence and Data Mining (CIDM)*, 2014.

George C Linderman and Stefan Steinerberger. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 1(2):313–332, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository (CoRR)*, abs/1907.11692, 2019.

Vivien Marx. Seeing data as t-SNE and UMAP do. *Nature Methods*, 21(6):930–933, 2024.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *Computer Research Repository (CoRR)*, abs/1802.03426, 2018.

Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6:100–108, 2025.

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium Series on Computational Intelligence*, pp. 159–166, 2015.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Isaac J Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39 (4):811–841, 1938.

Erich Schubert and Michael Gertz. Intrinsic t-stochastic neighbor embedding for visualization and outlier detection: A remedy against the curse of dimensionality? *International Conference on Similarity Search and Applications*, 2017.

Uri Shaham and Stefan Steinerberger. Stochastic neighbor embedding separates well-separated clusters. *Computing Research Repository (CoRR)*, abs/1702.02670, 2017.

Szymon Snoeck, Noah Bergam, and Nakul Verma. Compressibility barriers to neighborhood-preserving data visualizations. *Computing Research Repository (CoRR)*, abs/2508.07119, 2025.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. *International Conference on World Wide Web (WWW)*, pp. 287–297, 2016.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008.

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research (JMLR)*, 11(2), 2010.

Ben Weinkove. Stochastic neighborhood embedding and the gradient flow of relative entropy. *Computing Research Repository (CoRR)*, abs/2409.16963, 2024.

Zhirong Yang, Yuwei Chen, and Jukka Corander. t-SNE is not optimized to reveal clusters in data. *Computing Research Repository (CoRR)*, abs/2110.02573, 2021.