What will happen when we radically simplify t-SNE and UMAP visualization algorithms? Is it worth doing so?

Bartosz Minch[0000−0002−0122−1345], Radosław Łazarz[0000−0002−3151−9759] and Witold Dzwinel[0000−0001−8321−5928]

AGH University of Krakow, Poland
{minch,lazarz,dzwinel}@agh.edu.pl

Abstract. We investigate how the quality and computational complexity of the golden standards of high-dimensional data (HDD) visualisation - the t-SNE and UMAP algorithms - change with their successive simplifications. We show that by radically reducing the number of the utilised nearest neighbours, introducing binary distances between the samples, and simplifying the loss function, the resulting IVHD algorithm still reconstructs with sufficient precision both local and, particularly, global properties of HDD topology. Although inferior to its competitors for the most moderate data sizes (M<10^5 samples), IVHD appears many times faster than state-of-the-art algorithms and reveals its power for multi-million-element datasets for which baseline methods fail in a reasonable computational time.

Keywords: high-dimensional data · data embedding · kNN graph visualization · dimensionality reduction

1 Introduction

In recent years, the explosion of digital datasets resulted in new opportunities and challenges for various fields, such as machine learning, computer vision, bioinformatics, and social network analysis. One of the significant problems related to this data deluge is the high-dimensionality (HD) of the underlying objects, often represented by HD feature vectors with tens, hundreds, or even thousands of dimensions [6]. Their size can be especially burdensome for data analysis, causing increased computational and memory requirements, overfitting, and difficulties in visualisation or interpretation.

To address those issues, researchers developed dimensionality reduction (DR) methods that reduce the number of dimensions in the data while preserving the essential local and global topological properties. DR involves transforming the N-dimensional (N-D) dataset $X = \{x_i\}_{i=1,...M} \in \Re^N$ into its n-dimensional (n-D) representation $Y = \{y_i\}_{i=1,...M} \in \Re^n$, where $N >> n$, and M represents the number of N-D feature vectors x_i (or their corresponding n-D embeddings y_i). This transformation can be perceived as a lossy data compression, achieved by minimizing a loss function $E(|X - Y|)$, where $|.|$ measures the topological dissimilarity between X and Y.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23

In the world of unsupervised HDD embedding, t-SNE [12] and UMAP [7] are among the most widely adopted techniques. The former algorithm resembles the classical Multidimensional Scaling (MDS) [5], but instead of a simple loss function based on the cumulative L2 (or L1) discrepancies between distances in the source and target spaces, t-SNE compares the probability distributions of being a neighbour of each data vector using the Kullback-Leibler (K-L) divergence and summarizes them to calculate the actual embedding error. Those probability distributions reflect the neighbourhood of each data vector in the source and target spaces, with the highest probability assigned to the nearest neighbours (and its value decreasing rapidly in the case of the more distant elements).

On the other hand, UMAP corresponds to the Isomap [5] method, although it has a different conceptual basis than t-SNE and a distinct loss function for error minimization. Instead of calculating the full distance matrix or its Barnes-Hut approximation (as t-SNE does), UMAP focuses on the weights of the nearest neighbours of each data vector and a set of more distant samples. Nevertheless, there is a hidden relationship between t-SNE and UMAP. As demonstrated in [2], a generalisation of negative sampling allows the user to interpolate between embeddings produced by the two methods. Here we show that, additionally, t-SNE and UMAP can be further simplified to a frugal but still efficient approximation.

We have called this approximation the Interactive Visualization of High-dimensional Data — IVHD. It enables visualization of HDD structures in 2D or 3D Euclidean spaces by utilising their k-NN graph representations and the classical MDS loss functions. Moreover, it is assumed that the distances between nodes of the said graph follow the negative sampling principle, i.e. they are set to 0 for each node within the k-NN set — and 1 for m other randomly selected disconnected nodes. In practice, we observed [4, 3] that both k and m can be small, usually equal to 1 and 2, respectively. As demonstrated in [4], this concept allows visualising both complex networks (e.g., structured, random, scale-free, etc.) and high-dimensional data without changes in the base algorithm. In this paper, we highlight the following contributions:

- The IVHD method can be regarded as a unifying simplification of the t-SNE and UMAP algorithms, achieved through three approximation steps: (1) employing a simpler loss function, (2) utilizing binary distances, and (3) reducing the number of nearest neighbours used in embedding.
- We show that IVHD effectively preserves local and global properties of HDD in 2D embeddings, not only for large datasets but also for small ones, despite its simplicity. Furthermore, IVHD proves to be highly time-efficient when compared to the original methods.
- Additionally, we propose several novel improvements to IVHD (see Section 3).

2 Simplifying t-SNE and UMAP

Despite extensive research conducted in the field of HDD visualisation over the past years, new methods continue to emerge regularly [2]. In our research, we specifically investigate publicly available algorithms that have demonstrated superior performance in generating embeddings for datasets of varying sizes, including UMAP [7], IVHD [3],

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23

t-SNE, TriMAP, and PaCMAP [12]. Apart from t-SNE, these algorithms involve two key stages: (1) constructing a weighted k nearest neighbour graph and (2) performing an embedding procedure that involves defining a loss function and minimizing it [7]. In this study, we focus primarily on t-SNE and UMAP, which have been the foundation of recent research in this area, and aim to demonstrate that IVHD serves as a more parsimonious approximation for both methods.

2.1 Evaluation criteria

To verify the properties of the obtained simplifications, we use state-of-the-art quality assessment criteria [9] for unsupervised DR methods, measuring the preservation of the high-dimensional (HD) neighbourhood in the low-dimensional (LD) space. The general consensus is to use the average agreement rate between k-ary neighbourhoods in high and low dimensions. The rank of x_j with respect to x_i in a high-dimensional space is defined as $\rho_{ij} = \left|\{k : \delta_{ik} < \delta_{ij} \ \vee \ (\delta_{ik} = \delta_{ij} \ \wedge \ 1 \leq k < j \leq N)\}\right|$, where δ_{ij} is the distance between the i-th and j-th data point in HD (d_{ij} denotes an analogous distance in LD, respectively). Similarly, the rank of y_j relative to y_i in the low-dimensional space is equal to $r_{ij} = \left|\{k : d_{ik} < d_{ij} \ \vee \ (d_{ik} = d_{ij} \ \wedge \ 1 \leq k < j \leq N)\}\right|$. Let $\mathbf{v}_i^k$ and $\mathbf{n}_i^k$ represent the sets of nearest neighbours of x_i and y_i in the high-dimensional and low-dimensional space, with k denoting the number of those neighbours. Now, we define:

$$R_{NX}(k) = \frac{(N-1)\left(\frac{1}{kN}\sum_{i=1}^{N}\left|\mathbf{v}_i^k \cap \mathbf{n}_i^k\right|\right) - k}{N-1-k}, \quad G_{NN}(k) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|j \in \mathbf{n}_i^k\right| - \left|j \in \mathbf{v}_i^k\right|}{k}. \quad (1)$$

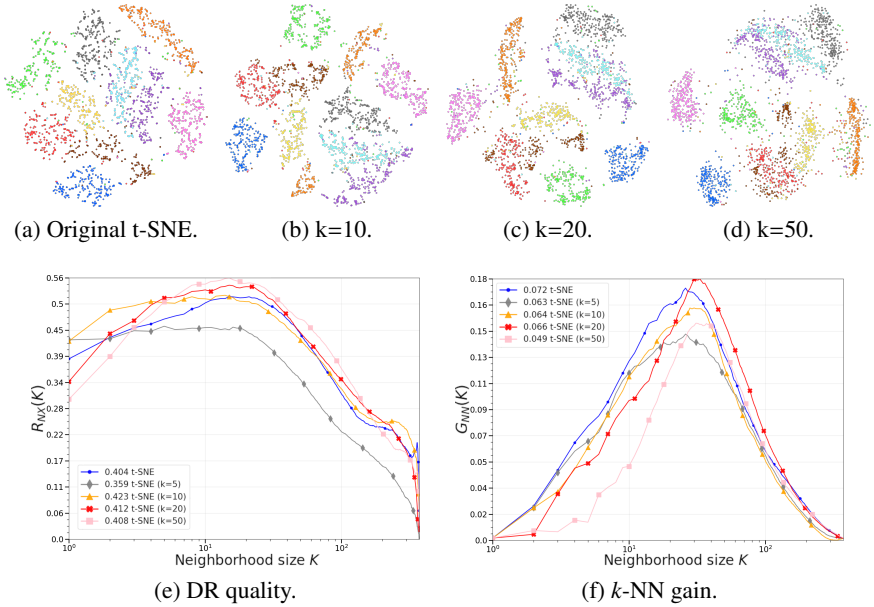
$R_{NX}(k)$ quantifies the quality improvement over a random embedding, while $G_{NN}(k)$, measures the average gain (or loss, if negative) considering neighbours of the same class, with a positive value indicating potentially better k-NN classification performance.

2.2 t-SNE with Euclidean and binary distances

We investigated whether t-SNE can be viewed as an embedding of an undirected graph, thereby enabling simplification to the IVHD framework. To explore this, we proposed a modified version of t-SNE that uses neighbourhood-limited Euclidean and binary distances instead of the standard probability matrix. This modification allowed us to parametrise t-SNE by the number of nearest neighbours (k) instead of the default perplexity (a critical parameter of t-SNE used to balance the local and global aspects). In this case, k determines the number of nearest data points considered when computing the probability distribution over the pairwise similarities in high-dimensional space. For the binary matrix variant, 1's were inserted to denote neighbours, and 0's otherwise. A similar approach was used for Euclidean distances (refer to [11]), but instead of 1's, the actual Euclidean distance was inserted.

As illustrated in Fig. 1 (and Fig. 5 in the supplementary materials [11]), for $k \in \{10, 20\}$, our variant of t-SNE achieves a DR quality comparable to the unmodified t-SNE when visualising a small subset (10%) of the MNIST dataset ($M = 7 \cdot 10^3$), as

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23



(a) Original t-SNE. (b) k=10. (c) k=20. (d) k=50.

(e) DR quality. (f) k-NN gain.

Fig. 1. The visualizations and metrics were obtained using a simplified t-SNE method on a 10% subset of the MNIST dataset, with binary distances instead of probabilities. The method was parametrised by the number of nearest neighbours k.

testified by the overlapping metrics curves. Furthermore, the more neighbours were used, the more the quality of visualization improved, as reflected by the higher reaching curves in both discussed graphs, resulting in better AUC values.

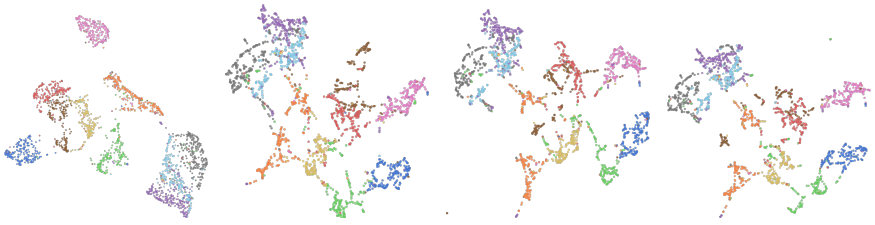
Additionally, it should be mentioned that simplifying the t-SNE method into an IVHD-based implementation also consisted of replacing the part of the algorithm that optimizes K-L divergence (a key aspect of DR methods that rely on neighbour embedding) with an optimisation of the MDS-like cost function. As a side effect, it led to obtaining a more streamlined and simplified IVHD method [3, 4, 10].

2.3 UMAP with a low negative sample rate and a small number of nearest neighbours

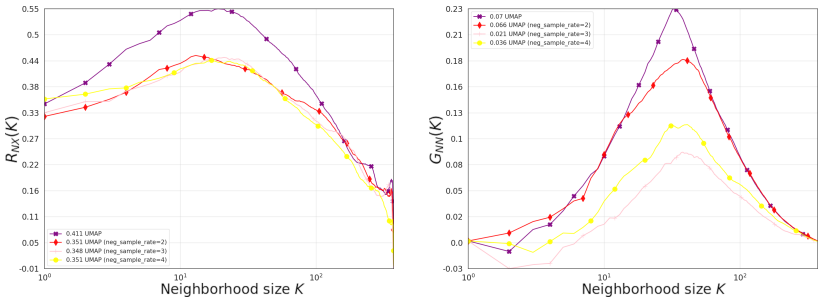
As mentioned in [2], there is a significant connection between negative sampling (NEG) and noise-contrastive estimation (NCE). UMAP, which uses NEG, can be seen as Neg-t-SNE [2], differing only in the implicit use of a less numerically stable similarity function. A key factor contributing to UMAP's success is its utilization of NEG to refine the Cauchy kernel and its cross-entropy loss function, which distinguishes it from how t-SNE assesses high-dimensional similarities. This refinement allows UMAP to generate more compact clusters and continuous connections between them, as demonstrated in [1]. Another perspective to consider is the similarity between UMAP and IVHD, which, similarly to MDS, aims to preserve pairwise distances or dissimilarities between high and low dimensional data points. To achieve this, UMAP constructs a

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23

distance matrix from the original HD data and then finds an LD embedding that minimises the difference between pairwise distances, while also preserving both local and global structures through the construction of a weighted nearest-neighbour graph. Subsequently, it optimizes that LD embedding to retain the aforementioned graph structure. By decreasing the rate of negative sampling and the number of nearest neighbours, we moved UMAP towards an IVHD-like simplification. In contrast to IVHD, UMAP is based on the idea of preserving the local structure of the data in an LD space, rather than just pairwise distances. This means that UMAP is still better adjusted to capture the non-linear relationships between data points.



(a) NSR=4, k=12 (default). (b) NSR=2, k=3. (c) NSR=3, k=3. (d) NSR=4, k=3.



(e) DR quality. (f) k-NN gain.

Fig. 2. UMAP visualizations and metrics were generated using different negative sampling rates and the lowest possible value of k (i.e. $k = 3$) that resulted in meaningful visualizations.

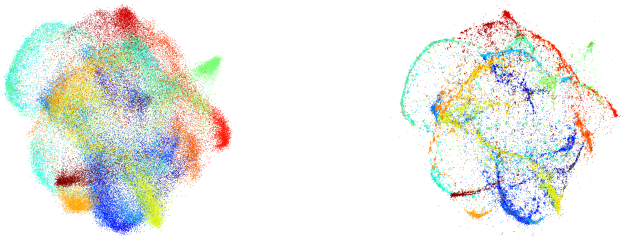
3 Improvements in the IVHD algorithm

IVHD, as described in [3, 10, 9], exceeds state-of-the-art DR algorithms in computational time by over tenfold in standard DR benchmark datasets [10]. The visualizations obtained are also proficient in reconstructing data separation in large, high-dimensional datasets [11]. Nevertheless, there is potential for improvement in terms of reducing the amount of noise generated between classes. To address this concern, we have developed the following improvements.

Reverse nearest neighbour (RNN) procedure. The query retrieves all points in a HD space that have a given point q as their nearest neighbour. The set of these points is called the influence set of q. It is important to note that a point p being one of the nearest neighbours k to q does not necessarily imply that p is also in q's *RkNN* set.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23

Manhattan norm employed in the final stages of the embedding procedure, providing a smoothing effect. Using this instead of the conventional Euclidean distance, outliers are drawn closer to the cluster centres. However, it is important to be cautious with the number of steps performed using the L1 metric, as excessive steps may cause the embedding to collapse towards the cluster centres, resulting in distortion of the global structure.



(a) The baseline IVHD. (b) With RNN and L1.

Fig. 3. A comparison of IVHD on the EMNIST dataset was conducted, employing the L1 norm and RNN mechanisms in the final steps of the embedding procedure.

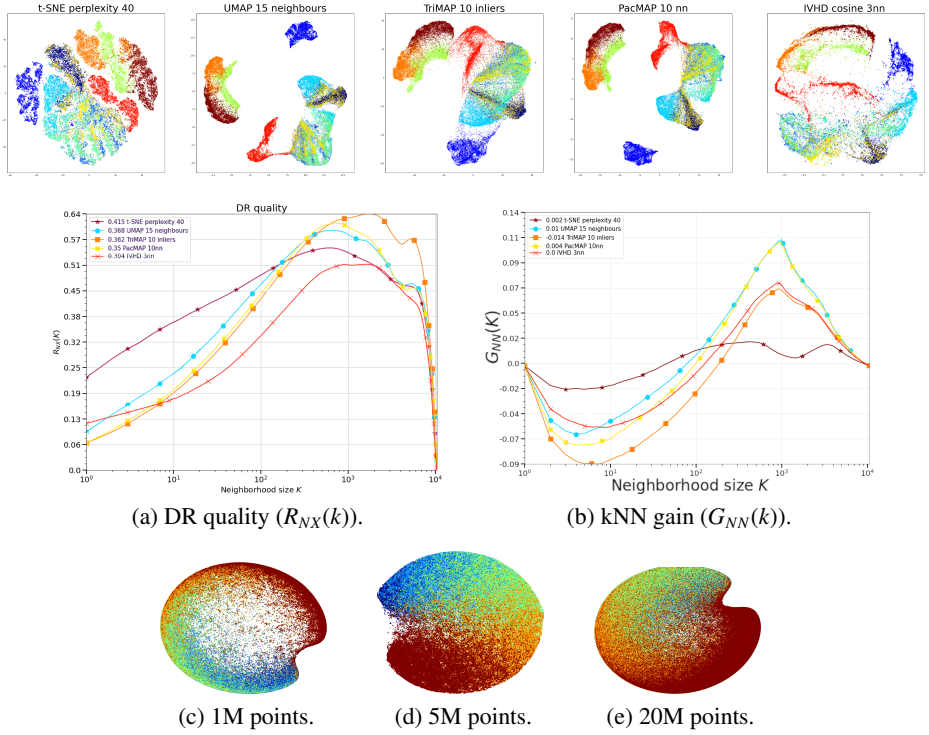
Fig. 3 illustrates the effect of incorporating both mechanisms to provide a generic approach for handling the noise remaining in the visualisation. Quality measurements demonstrate that the discussed upgrades to IVHD enhance the DR quality and the k-NN gain of the obtained embedding. In particular, a clear *suction effect* is observed, where most of the noise is moved from the interstitial space to the clusters. Importantly, the global structure of the visualization is preserved without distortion, as evidenced by the relative positioning of classes remaining unchanged.

The proposed improvements do not introduce significant computational overhead, as the helper graph for RNN is created concurrently (or retrieved from cache) together with the main graph, and interactions between a limited subset of nearest neighbours are calculated. The primary operation that incurs overhead is the search for reverse neighbours based on the two graphs, but the time taken by this procedure is negligible compared to the overall embedding time. It should be noted that all the improvements added to the IVHD method in this study were designed with the consideration of minimizing computational load, given the importance of efficient performance in HDD analysis.

4 Large-scale experiments

The primary benefit of IVHD lies in its ability to generate embeddings at a significantly faster rate than baseline methods, once the k-NN graph is stored in the disc cache and the computational time required for its generation can be disregarded. Consequently, users can perform detailed interactive analyses of the multi-scale data structure with a wide range of parameter values and stress function versions without the need to recalculate the said graph. In this regard, we present results obtained for the FMNIST (mid-sized) and the Amazon20M (large-sized) datasets as evidence of the efficiency and effectiveness of the IVHD (with $L1$ and RNN) method. Additional results are provided in the supplementary materials [11].

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23



(a) DR quality ($R_{NX}(k)$). (b) kNN gain ($G_{NN}(k)$).

(c) 1M points. (d) 5M points. (e) 20M points.

Fig. 4. Comparison of different DR methods employed on the FMNIST dataset. On the bottom: IVHD visualizations obtained for Amazon20M datasets.

In Figure 4, we observe that the IVHD applied to Fashion-MNIST forms separate groups of mostly elongated shapes. On the other hand, the MAP-family methods create rounded and clearly separated clusters.Additionally, in t-SNE, some classes are mixed and fragmented. In terms of DR quality, TriMap, UMAP, and PaCMAP are achieving the best results. IVHD surpasses t-SNE only when a large neighborhood is considered ($k > 1000$). Furthermore, IVHD-CUDA [10] was the only method capable of generating visualizations for the Amazon20M dataset in a reasonable time frame of 5 hours and 34 minutes. The generated visualizations clearly depict the separation of the five classes, which comprises book reviews from the Amazon platform. In contrast, other methods, including those implemented in both CPU and GPU environments (e.g. t-SNE CUDA, Anchor-tSNE [9]), encountered errors or did not generate visualizations even after 12 hours of calculations, leading to premature termination. Furthermore, Table 1 corroborates that IVHD stands out as the fastest method among the compared approaches.

	M	N	t-SNE	UMAP	TriMAP	PaCMAP	IVHD
EMNIST	784	103 600	558,67	50,25	123,84	79,21	**34,88**
REUTERS	30	804 409	7457,15	1226,91	1498,96	851,46	**190,47**

Table 1. A selection of timings (measured in seconds) acquired from various datasets with M representing the dataset dimensionality and N representing the number of samples.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23

5 Conclusions

We demonstrate that IVHD represents a radical simplification of both t-SNE and UMAP, offering a highly efficient platform for fast and interactive HDD visualization. Although it may produce slightly inferior embeddings compared to its competitors for small and moderate data sizes ($M << 10^5$ samples), IVHD remains a remarkably efficient algorithm. It accurately reconstructs both local and global data topology with precision, and its key advantage lies in its computational efficiency, enabling the visualization of large multi-million datasets within a reasonable time frame, where other baseline algorithms fail. We successfully verified the applicability of IVHD to the Amazon20M dataset, highlighting its unique ability to handle such large datasets with minimal resource utilization. Future research will be directed towards further improvements in IVHD, particularly in addressing the challenges of crowding and noise reduction.

Hardware All CPU implementations were executed in two environments, depending on the scale of the dataset processed. Mid-sized datasets were visualized on Macbook Pro 2.3 GHz 8-Core Intel Core i9, 16 GB 2667 MHz DDR4. Large-sized datasets were processed in GPU/CUDA remote server with Intel Xeon E5-2620 v3 CPU, 8GB GDDR5 NVidia GeForce GTX 1070 GPU, and 252 GB RAM. The source code was compiled using GCC-10.4 and CUDA Toolkit 11.2. Experiments were facilitated by the VisKit C++ library [8] developed by the first author of this work.

Acknowledgments The research presented in this paper was supported by funds allocated to the AGH University of Science and Technology by the Polish Ministry of Science and Higher Education. The authors utilized the PL-Grid Infrastructure and computing resources provided by ACK Cyfronet.

References

1. Böhm, J.N., Berens, P., Kobak, D.: Attraction-repulsion spectrum in neighbor embeddings. Journal of Machine Learning Research **23**(95), 1–32 (2022)
2. Damrich, S., Böhm, J.N., Hamprecht, F.A., Kobak, D.: Contrastive learning unifies *t*-sne and umap (2022), `https://arxiv.org/abs/2206.01816`
3. Dzwinel, W., Wcislo, R., Matwin, S.: 2-d embedding of large and high-dimensional data with minimal memory and computational time requirements. arXiv preprint arXiv:1902.01108 (2019)
4. Dzwinel, W., Wcislo, R., Strzoda, M.: ivga: Visualization of the network of historical events. In: Proceedings of the 1st International Conference on Internet of Things and Machine Learning '17. ACM (2017)
5. Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M.: Multidimensional scaling, sammon mapping, and isomap: Tutorial and survey (2020), `https://arxiv.org/abs/2009.08136`
6. Jia, W., Sun, M., Lian, J., Hou, S.: Feature dimensionality reduction: a review. Complex & Intelligent Systems **8** (01 2022)
7. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
8. Minch, B.: Viskit library. `https://gitlab.com/bminch/viskit`
9. Minch, B.: In search of the most efficient and memory-saving visualization of high dimensional data (2023)
10. Minch, B., Nowak, M.P., Wcislo, R., Dzwinel, W.: Gpu-embedding of knn-graph representing large and high-dimensional data. Computational Science – ICCS 2020 **12138**, 322 – 336 (2020)
11. Minch, B., Łazarz, R., Dzwinel, W.: Supplementary materials. `https://www.dropbox.com/s/s9wx5bz8wsq1zh3/ICCS_2023_Supplementary_Materials.pdf?dl=0`
12. Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y.: Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. Journal of Machine Learning Research **22**(201), 1–73 (2021)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_23