Bioinformatics, 2024, **40(7)**, btae471
https://doi.org/10.1093/bioinformatics/btae471
Advance Access Publication Date: 25 July 2024
Applications Note

OXFORD

Gene expression

Compound-SNE: comparative alignment of t-SNEs for multiple single-cell omics data visualization

Colin G. Cess[1] and Laleh Haghverdi [1,*]

[1]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (BIMSB-MDC), Berlin 10115, Germany

*Corresponding author. Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (BIMSB-MDC), Hannoversche Str. 28, Berlin 10115, Germany. E-mail: laleh.haghverdi@mdc-berlin.de (L.H.)

Associate Editor: Inanc Birol

Abstract

Summary: One of the first steps in single-cell omics data analysis is visualization, which allows researchers to see how well-separated cell-types are from each other. When visualizing multiple datasets at once, data integration/batch correction methods are used to merge the datasets. While needed for downstream analyses, these methods modify features space (e.g. gene expression)/PCA space in order to mix cell-types between batches as well as possible. This obscures sample-specific features and breaks down local embedding structures that can be seen when a sample is embedded alone. Therefore, in order to improve in visual comparisons between large numbers of samples (e.g. multiple patients, omic modalities, different time points), we introduce Compound-SNE, which performs what we term a soft alignment of samples in embedding space. We show that Compound-SNE is able to align cell-types in embedding space across samples, while preserving local embedding structures from when samples are embedded independently.

Availability and implementation: Python code for Compound-SNE is available for download at https://github.com/HaghverdiLab/Compound-SNE.

1 Introduction

Visualization of high-dimensional data is a key aspect when examining single-cell omics (epigenomics, transcriptomics, proteomics, etc.) data samples. Many different algorithms exist for embedding high-dimensional data into 2D space, though t-distributed Stochastic Neighbours Embedding (t-SNE) and uniform manifold approximation and projection (UMAP) remain the most common (Van der Maaten and Hinton 2008, McInnes *et al.* 2018). Besides visualizing a single sample, it is important to be able to visually compare multiple single-cell samples, e.g. scRNA-seq from different patient samples or data modalities such as paired scRNA-seq and scATAC-seq data (Kim *et al.* 2022) on the same sample or same patient, (multi-view data) before moving on to further analyses. If the dataset is complete (i.e. containing all cell states of interest) the reference dataset can be first embedded and other datasets projected onto it (Spitzer *et al.* 2015, Angerer *et al.* 2016, Hao *et al.* 2023). Otherwise, in current approaches, data integration is performed to merge samples together, which are then embedded all at once (Haghverdi *et al.* 2018, Korsunsky *et al.* 2019, Hao *et al.* 2021). While this does achieve a good alignment of different samples, data integration algorithms modify gene expression values in order to best mix samples together, leading, in embedding space, to the dissolution of unique local structures that are seen in original, unintegrated embeddings. Although data integration is still important for other analyses [e.g. such as cell type label transfer tasks (Mölbert and Haghverdi 2023)], we propose here an alternative method for visualizing multiple single-cell samples. Compound-SNE performs what we term a soft alignment, aiming to maximize the alignment of multiple embeddings while minimizing the local structural differences from the samples' independent embeddings. This is done in a two-step process: (i) alignment in PCA space via matrix transformation in order to align embedding initializations, and (ii) addition of a force term to the embedding algorithm, which pulls clusters of cells together based on annotations.

2 Alignment overview

The complete workflow of Compound-SNE consists of five steps as follows. Compound-SNE is designed to work with Scanpy (Wolf *et al.* 2018) formatting, taking in an AnnData object.

1) **Data processing:** Data should be processed via whatever method the user deems suitable and transformed into PCA (principal component analysis) space. Following Scanpy, this should be stored in the AnnData object as .obsm['X_pca']. While different samples and modalities may contain different number of original features, they should share the same number of features in PCA space, though data integration should not be performed here.

   In addition, for alignment, Compound-SNE requires cell annotations, ideally cell types, though other types of annotations are acceptable. If this is not available, Compound-SNE takes one sample as a reference,

Received: 26 February 2024; **Revised:** 10 July 2024; **Editorial Decision:** 17 July 2024; **Accepted:** 24 July 2024

performs k-means clustering, and, using the cells closest to each cluster centroid, identifies corresponding centroids in the other samples using mutual nearest neighbors (Haghverdi *et al.* 2018) in PCA space. We note that this in not preferred and that these clusters are only used for visual alignment and not any type of functional identification. Compound-SNE then integer encodes annotations. For the rest of the paper, we will refer to annotations as cell types.

2) **Reference selection:** Compound-SNE requires at least one sample to use as a reference. If not specified, Compound-SNE first chooses a primary reference as the sample with the most unique cell types. This sample is used for the primary alignment, as described in the following section. Then, if the primary does not contain all of the cell types, secondary references are chosen in order to complete the set of cells.
3) **Primary alignment:** Samples are first aligned in PCA space via matrix transformations in order to align cell-type centers. A cell-type center (only for shared cell types) x components matrix is found for each sample, which is then aligned to the primary reference, after scaling, via a Procrustes transformation (Schönemann 1966), which scales and rotates a matrix to minimize the sum of squared errors from a reference matrix. The obtained transformation matrix is used to transform the full sample. This impacts the embedding initialization, but has no further impact on the embedding process.
4) **Embedding initialization:** As Kobak and Linderman (2021) shows, initializing a nonlinear embedding optimization with PCA components enhances preservation of global structures. The authors also report that both t-SNE and UMAP equally preserve global structures when using the same initialization. We therefore use the first two components of the transformed PCA space in order to initialize the embedding for each sample.
5) **Alignment via forces:** To obtain better alignment between samples with minimal disturbance to local embedding structure, we include an additional force term to the embedding process that pulls the centers of cell type clusters (that may deviate from the primary alignment in the process of t-SNE iterations) together for each sample. We first embed the reference sample as normal, then find the centers of each type in embedding space. When embedding the remaining samples, during each embedding step, cell type centers are found and the distance between embedding sample centers and reference centers is found, with the goal of minimizing these distances. The total loss function thus becomes

$$L_{total} = L_{tsne} + \lambda_i \sum_{i=1}^{K} d_i$$

$$d_i = (Y_{r,center_i} - Y_{s,center_i})^2$$

where L_{tsne} is the standard t-SNE cost function (Van der Maaten and Hinton 2008), d_i is the squared distance between embedded reference centers, $Y_{r,center_i}$, and embedded sample centers, $Y_{s,center_i}$, for cell type i, and K is the number of shared clusters between the reference and the sample data. We expand upon the relation between the alignment force exertion and minimization of the loss function in the Supplementary Methods.

For practical implementation, we take advantage of the computational speed of the openTSNE (Poličar *et al.* 2024) Python library. Compound-SNE alternates between a t-SNE iteration, via openTSNE, and minimizing the distance between cell-type clusters.

Because not all samples may contain every cell type, as described above, the primary reference is chosen as the one with the most unique type. We then identify secondary references, using the minimum needed to create a set containing all of the present cell types. Secondary references are then aligned sequentially to the primary, using their embeddings to obtain embedding centers of remaining types. This creates a complete reference of embedding centers for each cell type present across all samples.

3 Application

We apply Compound-SNE to datasets consisting of multiple patients and modalities, demonstrating its utility for comparing different but related datasets. One dataset consists of bone marrow samples from six healthy patients, containing both gene expression and surface markers (Triana *et al.* 2021). The second dataset consists of gene expression and ATAC-seq data for kidney samples from the same patient (Muto *et al.* 2021). A subset of alignments is shown in Fig. 1, with full alignments in Supplementary Figs S1 and S2. The third dataset consists of gene expression of bone marrow hematopoietic cells for several time-points following inflammatory stimulation (Bouman *et al.* 2024) (shown in Supplementary Fig. S3).

In Fig. 1a, using gene expression of patient B6 as a reference, we show that Compound-SNE can be used to align gene expression for several patients. The first column shows the original, independent embeddings for each sample. The second columns shows embedding following the primary alignment and the third column shows embedding with the additional force term. The final two columns shows embedding following data integration using Harmony and Seurat, as a comparison to our method. Visually, we see that even using only the primary alignment offers a reasonable improvement over the independent embeddings, with the full alignment providing a much greater visual alignment. Notably, the full alignment yields embeddings that retain much of the cluster shapes that are seen in the independent embeddings. The two integration methods, while clearly aligning all of the samples, visually erase much of the structures unique to each patient in the independent embeddings. This is because cells are forced to mix well between batches.

In Fig. 1c, we align scRNA and scATAC samples from the same patient. While in comparison to Fig. 1a, where the independent embeddings look somewhat comparable between patients (as well as between scRNA and surface markers in Supplementary Fig. S1), the embeddings for scRNA and scATAC look very different from each other initially, obscuring comparison. Primary alignment achieves a modest improvement, while the full alignment yields a much stronger improvement while preserving original cluster shapes. We were unable to integrate the two modalities using Harmony, while Seurat was able to integrate them, again at the cost of dissolving structures present in the independent embeddings.

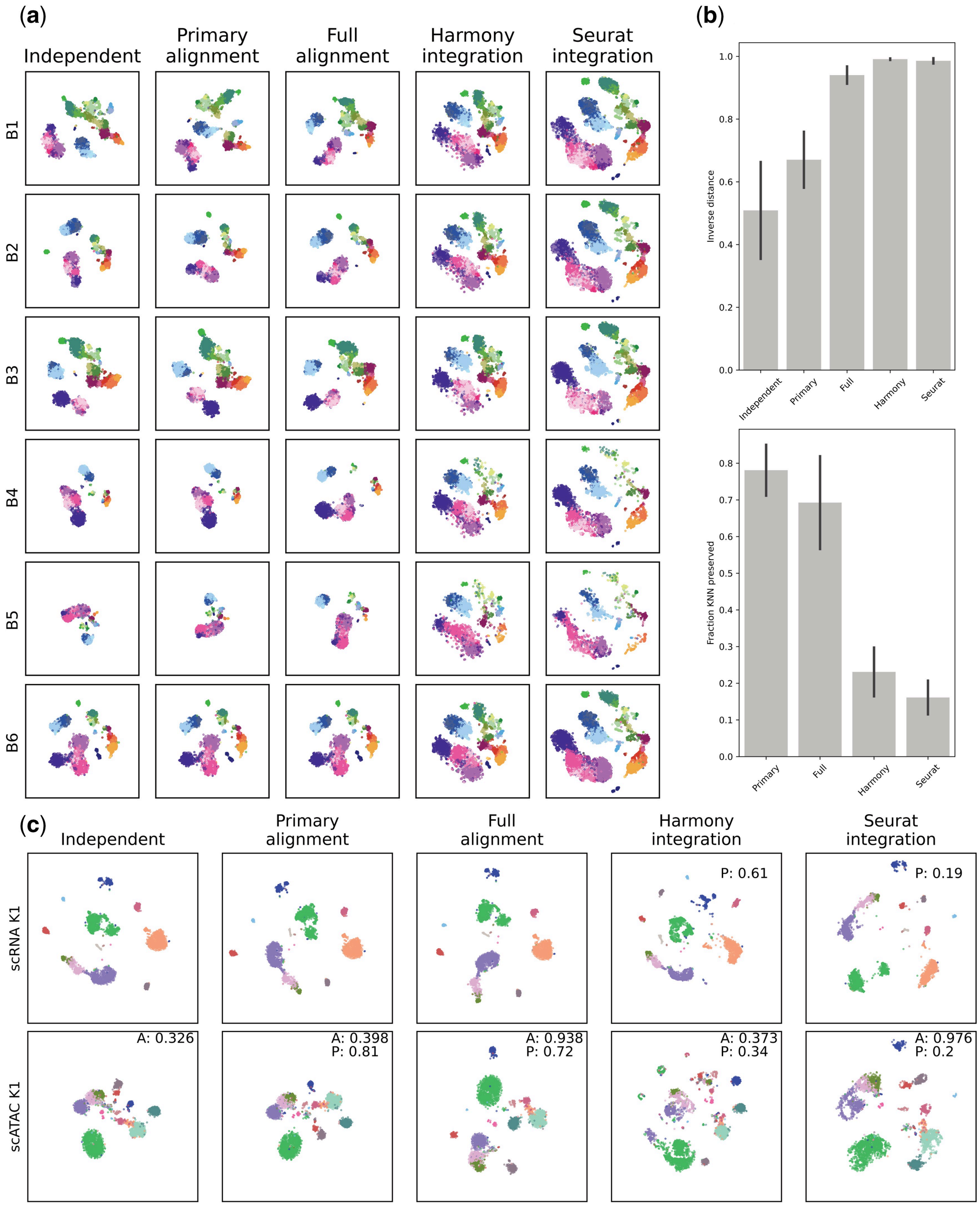

Figure 1. (a) Embeddings for the six patients from the bone marrow dataset, using B6 as the primary reference. Each row corresponds to a different alignment/integration method. All embeddings are on the same spatial scale. (b) Metrics for the embeddings shown in A. Means with error bars for standard deviation. Top: structure preservation, calculated as the fraction of KNN for each point preserved from the Independent embeddings. Bottom: alignment of the embeddings as the distance between normalized cell type centers. (c) Alignment/integration of scRNA and scATAC samples for K1 of the kidney dataset. Alignment scores of scATAC to scRNA are shown on each scATAC subplot, labeled as A. Structure preservation scores, labeled as P, for scATAC are shown on the subplots, excluding the Independent embedding. This score is also shown for integration methods on scRNA. All embeddings coordinates are on the same scale. Cell-type legends for (a) and (c) are shown in Supplementary Fig. S8.

3.1 Comparison statistics and evaluations

Beyond a visual comparison of embeddings, we calculate several metrics to compare how well-aligned embeddings are to each other and how well embedding structures are preserved between aligned embeddings and the original embeddings.

1) **Alignment score:** Beyond visually comparing embeddings, we calculate a metric to determine how well-aligned samples are. In embedding space, we find the centers of each cell-type for each sample and take the sum of squared errors between points. This value, d, is then transformed via $1/(1 + d)$ so that a value closer to 1 indicates a better alignment. We see that (Fig. 1b, top), as we progress from independent embeddings to aligned initializations to aligned with center-based force, we get better alignment, which is consistent with the visual results. We do see that data integration methods Harmony and Seurat yield the best alignment between samples, which is expected based on the nature of data

integration. Alignment scores between scRNA and scATAC for patient K1 are shown directly on the plots of Fig. 1c.

2) **Locality preservation:** While data integration yields the best alignment between samples, we can visually see that this is at the cost of the original embedding structure (Fig. 1b, bottom). To determine the preservation of local structures present in each embedding, we calculate the k nearest neighbors for each cell in the independent embedding and compare it to the nearest neighbors in each alignment, taking the fraction shared as a metric of structure preservation. We see that the primary alignment obtains the best preservation of original structure, with alignment with center-forces performing only slightly worse. Data integration, on the other hand, greatly disrupts these local structures. We therefore see that there is a trade-off between structure preservation and sample alignment. Preservation scores for scRNA and scATAC for patient K1 are shown directly on the plots of Fig. 1c.
3) **Alignment of data views with highly variable sizes (cell numbers):** Furthermore, to demonstrate the alignment of samples with highly different cell densities, we randomly subsample bone marrow B2 to 696 cells (1/10 of the cells) and align it with the full sample for B1 (9751 cells) (Supplementary Fig. S4). We see that this still achieves a nice visual alignment.
4) **Computational efficiency:** In Supplementary Fig. S5, we compare the runtime for each samples when embedded independently and embedded with alignment forces. We find that, with a couple of outliers in either direction, the addition of alignment forces does not impact runtime.
5) **Clustering for cell annotations:** We mentioned that Compound-SNE is able to generate noncell-type-specific annotations for the sake of performing alignment. Applying the full alignment to these generated annotations for the bone-marrow scRNA samples is shown in Supplementary Fig. S6, which shows a comparable alignment, in this case, to using the original cell-type annotations.

4 Conclusion

With Compound-SNE, we demonstrate how we can perform a soft alignment of embeddings for single-cell samples from different patients and modalities. This aids a visual comparison between many samples, with minimal disturbance to the unique sample structures seen when embedding samples independently.

When using Compound-SNE, the usual limitations in interpreting nonparametric data embedding (like standard t-SNE) should be respected (Chari and Pachter 2023). Whereas comparison of the overall structure, clusters composition and features activities (e.g. gene expression) across the map are correct and useful, over-interpretations such as comparison of cell densities over the maps should be avoided.

Author contributions

Colin G. Cess designed the algorithm, developed the software and performed data analysis. Laleh Haghverdi conceptualized the project. Colin G. Cess and Laleh Haghverdi wrote the manuscript.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) grant for “junior consortia in systems medicine” [01ZX1911B].

Data availability

Compound-SNE is available on Github (https://github.com/HaghverdiLab/Compound-SNE). The bone marrow dataset was downloaded from https://figshare.com/articles/dataset/Expression_of_97_surface_markers_and_462_mRNAs_in_70017_cells_from_healthy_young_healthy_old_and_leukemic_human_bone_marrow/13397651. The kidney dataset was downloaded from https://cellxgene.cziscience.com/collections/9b02383a-9358-4f0f-9795-a891ec523bcc.

References

Angerer P, Haghverdi L, Büttner M *et al.* Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 2016;**32**:1241–3.

Bouman BJ, Demerdash Y, Sood S *et al.* Single-cell time series analysis reveals the dynamics of HSPC response to inflammation. *Life Sci Alliance* 2024;**7**:e202302309.

Chari T, Pachter L. The specious art of single-cell genomics. *PLoS Comput Biol* 2023;**19**:e1011288.

Haghverdi L, Lun AT, Morgan MD *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87.e29.

Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2023;**42**:1–12.

Kim SH, Marinov GK, Bagdatli ST *et al.* Simultaneous single-cell profiling of the transcriptome and accessible chromatin using share-seq. In: *Chromatin Accessibility: Methods and Protocols*. New York, NY, US: Springer, 2022, 187–230.

Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol* 2021;**39**:156–7.

Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96.

McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426, 2018, preprint: not peer reviewed.

Mölbert C, Haghverdi L. Adjustments to the reference dataset design improve cell type label transfer. *Front Bioinform* 2023;**3**:1150099.

Muto Y, Wilson PC, Ledru N *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat Commun* 2021;**12**:2190.

Poličar PG, Stražar M, Zupan B. Opentsne: a modular python library for t-SNE dimensionality reduction and embedding. *J Stat Soft* 2024;**109**:1–30.

Schönemann PH. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 1966;**31**:1–10.

Spitzer MH, Gherardini PF, Fragiadakis GK *et al.* An interactive reference framework for modeling a dynamic immune system. *Science* 2015;**349**:1259425.

Triana S, Vonficht D, Jopp-Saile L *et al*. Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. *Nat Immunol* 2021; **22**:1577–89.

Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.

Wolf FA, Angerer P, Theis FJ. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15.

Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae471
Applications Note