Projection of high-dimensional cytometry data using regularised autoencoders

David Novak[1,2], Sofie van Gassen[1,2], Yvan Saeys[1,2]

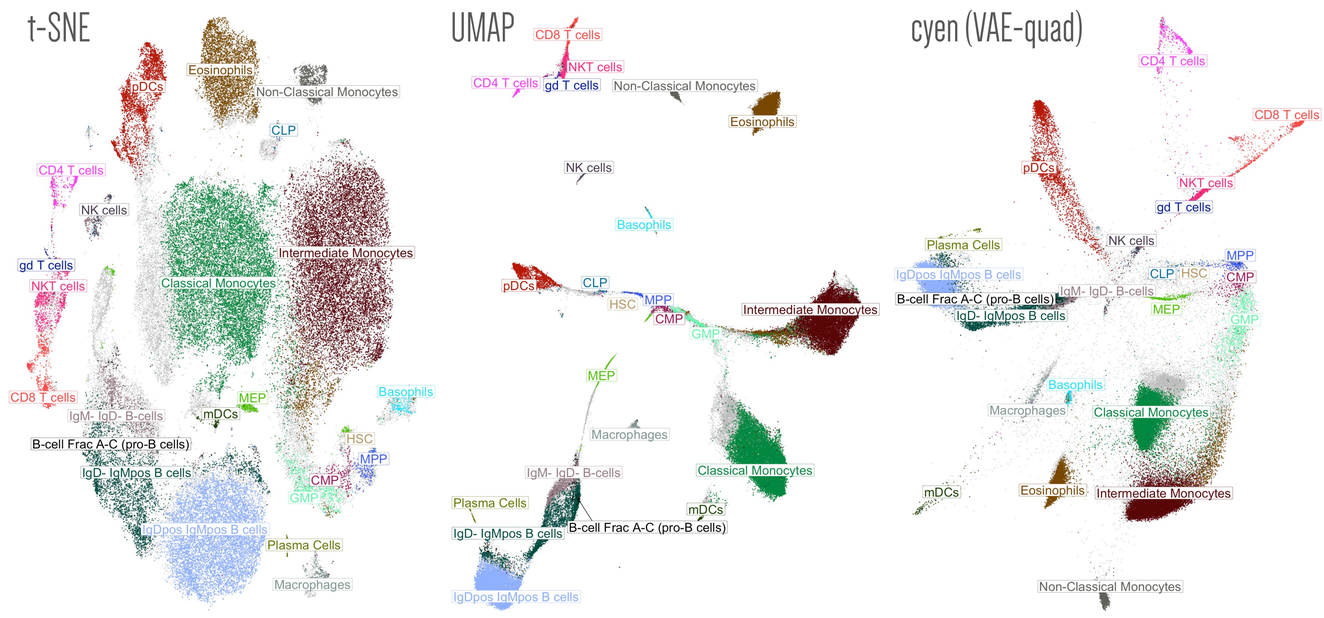
1 Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium.
2 Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium.

BACKGROUND

Variational autoencoders (VAEs), a type of artificial neural networks, have been used in visualisation (*ivis*, *scvis*) and clustering (*SAUCIE*, *MoE-Sim-VAE*, *VAE-SNE*). Conveniently, VAEs can **combine training objectives to preserve relationships between points at different scales**, can be made **robust to noise** and are **generative models**, allowing for imputation by sampling new data from learned distributions.

A new tool developed for cytometry data

cyen is a new model for learning lower-dimensional or denoised embeddings of biological data. This is done for **visualisation via dimension reduction (DR)**, **imputation** and **downstream data analysis (clustering, classification)**. Here, we showcase the DR component.



↑ 2-dimensional projections of the 39-dimensional Samusik CyTOF dataset [4] with default settings for t-SNE (R package `Rtsne`), UMAP (R package `uwot`) and a post-smoothed VAE-quad projection (VAE with quartet loss to preserve global relationships) from *cyen*.

RESULTS

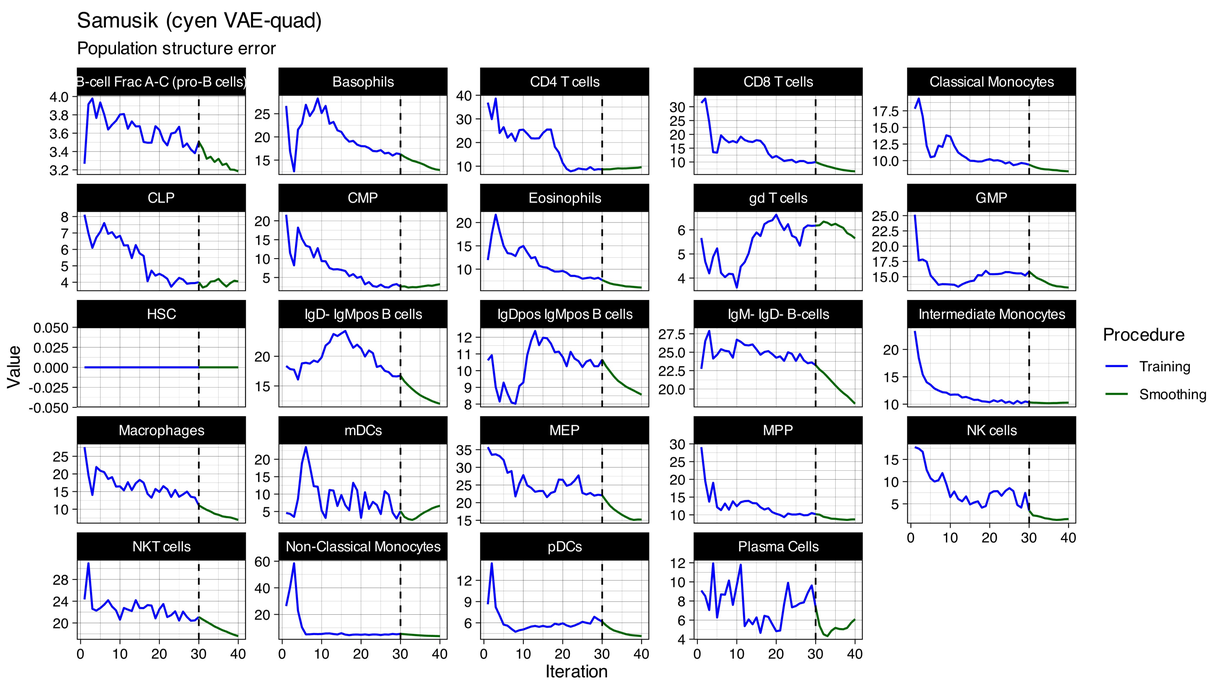
LOCAL STRUCTURE PRESERVATION

method \ dataset	Samusik RAW	Samusik POST-SMOOTHED	Levine32 RAW	Levine32 POST-SMOOTHED	Nilsson RAW	Nilsson POST-SMOOTHED
VAE	0.245 ± 0.008	0.257 ± 0.007	0.223 ± 0.005	0.254 ± 0.005	0.301 ± 0.046	0.315 ± 0.046
VAE-tri	0.254 ± 0.005	0.265 ± 0.005	0.228 ± 0.003	0.259 ± 0.005	0.306 ± 0.019	0.322 ± 0.019
VAE-quad	0.267 ± 0.003	**0.275 ± 0.004**	0.238 ± 0.002	0.269 ± 0.004	0.339 ± 0.002	0.352 ± 0.003
t-SNE		0.270 ± 0.008		**0.321 ± 0.007**		**0.386 ± 0.006**
UMAP		0.269 ± 0.001		0.274 ± 0.004		0.356 ± 0.004

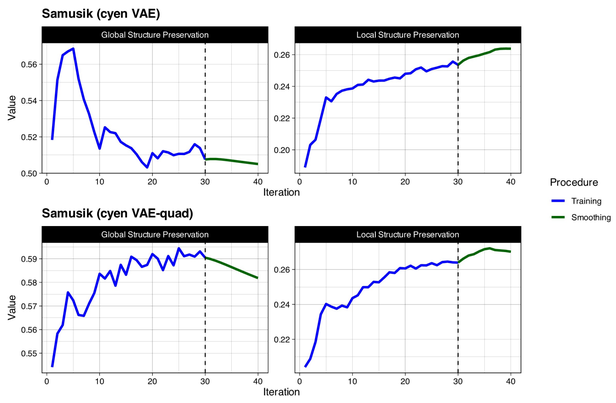
GLOBAL STRUCTURE PRESERVATION

method \ dataset	Samusik RAW	Samusik POST-SMOOTHED	Levine32 RAW	Levine32 POST-SMOOTHED	Nilsson RAW	Nilsson POST-SMOOTHED
VAE	0.503 ± 0.025	0.499 ± 0.024	0.491 ± 0.22	0.485 ± 0.025	0.611 ± 0.069	0.610 ± 0.067
VAE-tri	0.542 ± 0.021	0.537 ± 0.021	0.477 ± 0.33	0.470 ± 0.030	0.619 ± 0.029	0.617 ± 0.029
VAE-quad	**0.607 ± 0.013**	0.598 ± 0.012	**0.578 ± 0.009**	0.563 ± 0.008	**0.720 ± 0.007**	0.713 ± 0.007
t-SNE		0.335± 0.021		0.303 ± 0.022		0.352 ± 0.039
UMAP		0.492 ± 0.004		0.418 ± 0.007		0.553 ± 0.001

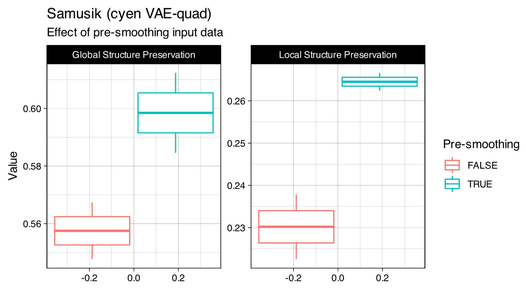
↑ Scoring for 3 *cyen* set-ups producing a 2-dimensional projection of 3 benchmark datasets [4], compared to t-SNE and UMAP. Each method was run 5 times with different random seeds (mean and standard deviation are shown). All *cyen* set-ups use a Gaussian-mixture latent prior and pre-smoothing for input data; effect of post-smoothing on *cyen* projections is also shown.



↑ Population structure error shows how well different populations are preserved with regard to the identity of cells in their near neighbourhood. Change with additional training and post-smoothing iterations for the *cyen* VAE-quad model is plotted here. We can see the model does not optimise for better structure for each population in each iteration, but effectively makes trade-offs between them.



↑ With a standard *cyen* VAE model we see local score increasing with training, but global score suffering. The addition of the quartet loss term (VAE-quad) prevents this trade-off and produces a better embedding.



↑ Pre-smoothing (denoising) input data before training the *cyen* VAE-quad model shows better structure preservation (5 runs with different random seeds are shown).

METHODS

Advanced benchmarking framework for cytometry DR

Benchmarking DR in cytometry is hard due to high cell counts. We use scalable approximations of **structure preservation** metrics (*Area Under RNX curve*) [1], extend them to **assess both local and global structure** separately and modify another metric (*Neighbourhood Proportion Error*) [2] to compute 'population structure error', scoring **reconstruction of labelled/clustered populations**, i.e. how faithful the embedding of each cell group is.

Our models can use **combinations of different loss functions (objectives)**.

Quartet loss ensures layout is meaningful at multiple scales

This setting (*VAE-quad*) is adapted from recent work on quartet-based stochastic embedding [3]. It preserves global relationships among cells.

t-SNE loss preserves small local neighbourhoods

The objective from *t-SNE* (in *VAE-sne*) preserves small neighbourhoods of points by minimising divergence between distribution fitted to nearby points in original data and in the embedding.

Triplet loss prevents intrusions and extrusions in close neighbourhoods

The triplet loss adopted from *ivis* trains a Siamese neural network (*VAE-tri* in *cyen*) to penalise faraway points from entering a close neighbourhood of a vantage point and vice versa.

Smoothing to denoise high-dimensional data

We boost performance by first **denoising** (*pre-smoothing*) input data: moving each point closer to average coordinates of its close neighbours. Smoothing can also be used on embeddings after training (*post-smoothing*).

DISCUSSION

Our **evaluation framework** allows for optimising DR methods for cytometry datasets. This allows us to experiment further with **combinations of loss functions** and **pre-training strategies** (changing objectives between training epochs) in *cyen*, with the aim to surpass t-SNE and UMAP in both local and global structure preservation. Additionally, we are working toward embeddings that yield better results in automated partitioning of cell populations and classification of cell types than the original high-dimensional sparse data. This work generalises well to other **noisy high-dimensional data in biology**.

[1] de Bodt, C. et al. (2020). DOI: 10.1109/TNNLS.2020.3042807
[2] Konstorum, A. et al. (2018). DOI: 10.1101/273862v2
[3] Lambert, P. et al. (2021). DOI: 10.1101/273862v2
[4] Weber, L. and Robinson, M. (2016). DOI: 10.1002/cyto.a.23030

This research is supported by FWO Strategic Basic research grant 1S40421N .

VIB-UGENT CENTER FOR INFLAMMATION RESEARCH
GHENT UNIVERSITY
Research Foundation Flanders
Opening new horizons
SCIENCE MEETS LIFE