ViScoreR: label-based evaluation of dimensionality reduction by detecting local distortions

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

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

DOWNLOAD AT
github.com/saeyslab/ViScoreR

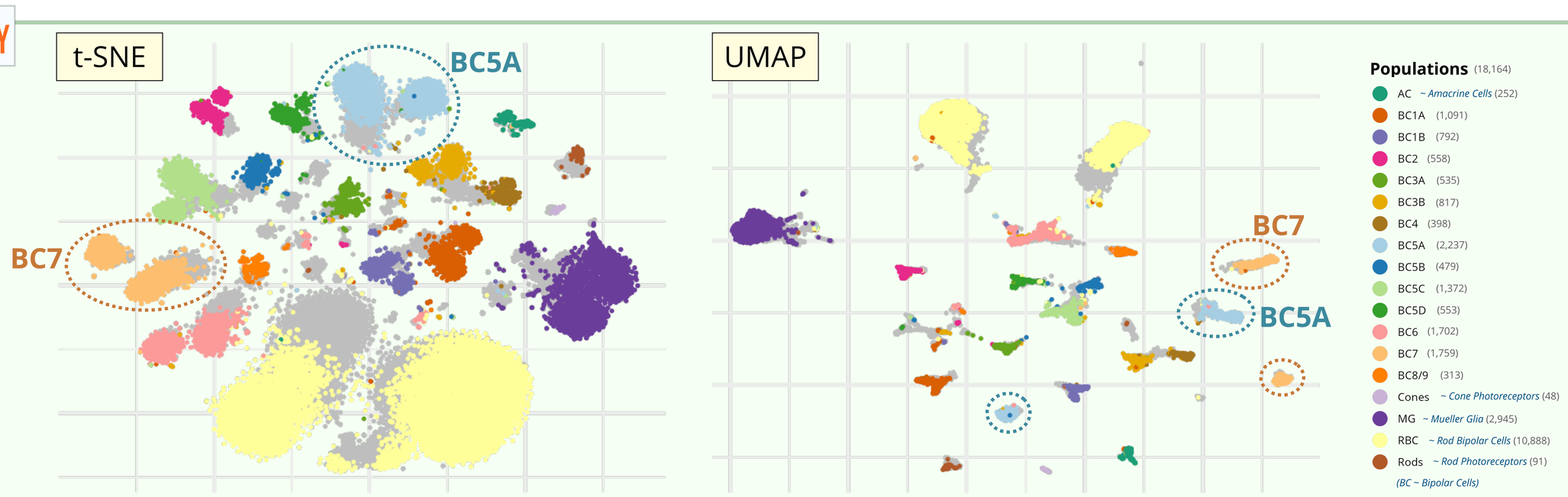
Introduction & methods

ViScoreR evaluates dimensionality reduction (DR) of single-cell data by comparing **neighbourhood relations of labelled cell populations** in the high-dimensional (HD) input data and the low-dimensional (LD) embedding, and between different LD embeddings. *Each DR tool introduces artifacts. ViScoreR is a diagnostic tool that identifies them easily, so as to prevent faulty reasoning about data based on a misleading embedding.*

xNPE[1] (population shape error): for each population (reference), how does the distribution of same-vs-differently labelled cells in the reference's neighbourhood change, as we go from HD to LD?
For comparing shape distortions in different LD embeddings.

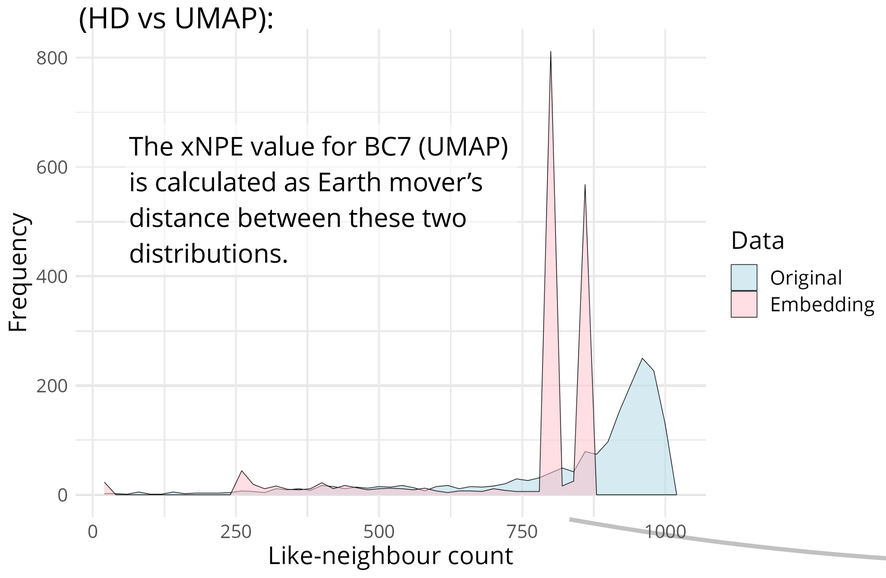
NCP (neighbourhood composition plots): which cell populations are represented in the near neighbourhood of a given population (reference), in HD data or in any given LD embedding of it?
For identifying sources of positional distortion.

Case study

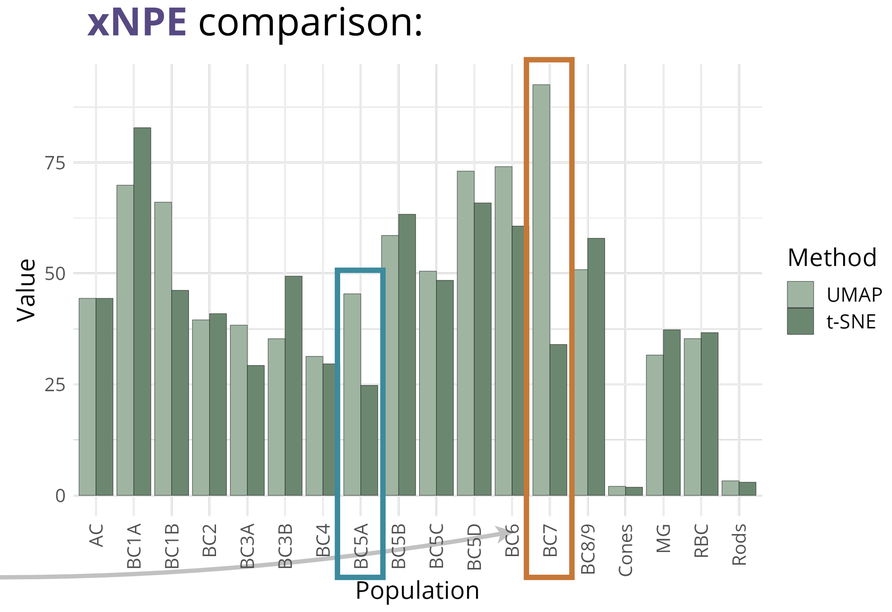


This dataset[2] has a known batch effect, leading to many populations being split into two clusters. In the case of **BC5A** and **BC7** cells, UMAP places the respective cluster pairs far apart, as opposed to t-SNE. We show that the **xNPE** score for UMAP is, accordingly, much higher than that of *t*-SNE for these populations. We use the **NCP** to see how the **BC7** is positioned with respect to other populations in the original data and in each embedding, revealing the nature of positional distortion in the embeddings.

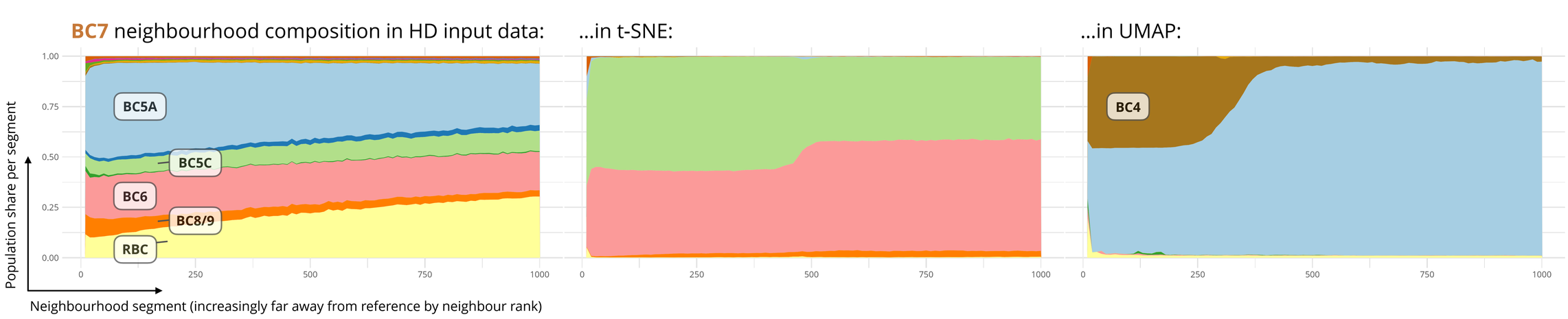
BC7 distribution of same-population neighbour counts (HD vs UMAP):



xNPE comparison:



BC7 neighbourhood composition in HD input data: ...in t-SNE: ...in UMAP:



↑ These plots let the user identify **what kind of positional distortion is occurring**. Considering BC7 as reference, t-SNE inflates the proximity of BC5C and BC6 cells (relative to the true neighbourhood profile in the original data), and fails to capture the proximity of BC5A cells. The UMAP embedding suffers from an intrusion of BC4 cells into the close neighbourhood of the reference and fails to capture many other populations in the close neighbourhood.

References

1. **xNPE (Extended Neighbourhood-Proportion-Error) is based on Neighbourhood Proportion Error:** Konstorum A et al. (2018) Comparative Analysis of Linear and Nonlinear Dimension Reduction Techniques on Mass Cytometry Data. bioRxiv 273862.
2. **Analysed data from:** Shekhar K et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 166(5), 1308-1323.

VIB-UGENT CENTER FOR INFLAMMATION RESEARCH
GHENT UNIVERSITY

fwo Opening new horizons
This work is funded by the FWO (Fonds Wetenschappelijk Onderzoek) and conducted as part of the Strategic Basic research project 1S40421N.

SCIENCE MEETS LIFE