**Poličar, Stražar & Zupan: Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification**

**RESPONSE TO THE REVIEWERS**

Dear Editor and the Reviewers,

we thank for the comprehensive feedback that helped to improve the manuscript, sufficient time to implement the changes, and the permission to submit a revised version.

All of the reviewer remarks are addressed below with reference to the corresponding changes in the manuscript. In case of overlapping remarks from different reviewers, we refer the reader to the relevant enumerated answer. A differential document is provided to facilitate the comparison with the initial submission. The changes to the data, code and the instructions to reproduce all the experiments are available in the repository: `https://github.com/biolab/tsne-embedding`.

Kind regards,

Pavlin Poličar
Martin Stražar
Blaž Zupan

University of Ljubljana, Slovenia

## Reviewer #1

> 1. I am missing more comparisons to the standard t-SNE, i.e., more examples where the proposed method removes the batch effect. The results section is focused on demonstrating that the approach appropriately embeds the second dataset to the reference dataset (in terms of cell types), however, it gives few examples of the batch effect and how the proposed method resolves it. More comparison to the standard t-SNE would be welcome.

The batch-effect related to embedding multiple datasets is highly abundant. To reinforce this claim, we have added two additional experiments with independent datasets (...) in the accompanying notebooks. Also in these cases, the standard t-SNE would be confounded by data source of origin.

We have added a paragraph to Section 3.4, pointing the reader to the relevant material and findings.

> 2. Also, the community still largely uses PCA in order to visualize high dimensional data, therefore, the comparison to PCA would be interesting to see as well - this should be done very easily since the proposed method starts from the PCA-based representation of the data.

Since this comparison has been made several times in the literature (...), we opted to omit this comparison in favour of brevity. In molecular biology for example, PCA remains common for analysis of small-sample datasets with tens to hundreds of samples (e.g. bulk RNA sequencing), while t-SNE is essential for single-cell studies with tens of thousands to millions of samples.

The two methods are based on different practical objectives. Note that the objective of PCA is to find an optimal approximation to the data, *limited to* an orthogonal subspace of the input data. The linearity of the transformation additionally limits the capacity of 2D embeddings, typically used for visualization. While often useful to visualize smaller-to-medium sized datasets, the low-rank subspace constraint does not work well with increasing number of dimensions.

3. As far as I understand, the results are not symmetric in terms of which of the two datasets is a reference and which one is embedded? How should one decide which dataset is better used as a reference? Furthermore, how should more than two datasets be analyzed with the proposed approach?

We agree with the Reviewer's observation and have explicitly mentioned this property in the discussion (Section 3.4). From a theoretical standpoint, the choice of referenence is largely arbitrary. In practice, however, one would start with previously published reference lanscapes and embed own, new (secondary) datasets in order to identify cell types.

4. It would be interesting to report what are time and memory requirements of the proposed method. More specifically, what is the overhead of the proposed method in comparison to the standard t-SNE?

report the time

The embedding of each cell independently essentially reduced the complexity of embedding the secondary dataset to linear.

## Reviewer #2

5. Can you validate the results also *quantitatively*? Visualizations are nice, but numbers may be able to provide complementary information that the eye does not see.

The quantitative estimates of erroneus placed were estimated for each of the three reference/secondary data set pairs. Additionally, we have added two new experiments with cells from new studies, for which the error estimates are reported as well.

Empirically, the method incorrect placement ranges from 1-5 % of the cells, assuming ground truth annotation.

6. Can you *prove* anything about the results except that it is the result of an optimization?

The results in research of t-SNE and related non-linear methods from visualization are by and large statistical, empirical and subjective. The key property of our method formulation, which can be proved trivially, is the independence of embedding each secondary data point, which in turn leads to batch-effect removal.

(see what happens with randomly-permuted cell profiles)

7. In the abstract, I found the formulation "one data point at a time" misleading. I was expecting order-dependent effects, where you in fact meant that the optimizations are computed for each data point of the secondary dataset individually (and thus are in fact "embarassingly parallel").

We thank the reviever for this careful observation. The the text was revised at relevant places (Section 1) to explicitly state that data points are embedded independently of one another and of order.

8. Fig. 7, in particular as a comparison with a simpler method, does not look great from my point of view. I fail to see why the newly proposed method is better than the simpler method. Maybe you can elaborate a bit more on that.

As stated in Section 3.4, the placement od data points with t-SNE embedding is subject to optimization rather than interpolation (as with nearest neighbours approach). The latter is guarateed to place the secondary data points into existing clusters and perhaps prone to overclustering. With t-SNE embedding, this effect is less pronounced and data points unrelated to the reference can be identified as outliers. We have added a sentence to elucidate this difference between the methods.

(Perhaps this is shown with the permutation test.)

## Reviewer #3

> 9. no quantitative statistics are reported regarding the accuracy of the proposed algorithm in correctly classifying each cell in the correct type. Please consider adding this information

The concern was addressed in Reviewer Remark 5.

> 10. as the authors pointed out, one limitation of the proposed approach is that the reference dataset should already contain all possible cell types that are present in the other datasets. Would it be possible to label as a outlier any novel cell that does not fit within the cell types present in the reference dataset? Please elaborate

The concern was addressed in Reviewer Remark 3.