DeepSCore, a multi-language multi-omics framework for single-cell automatic label transfer

Pablo Iáñez Picazo & Elisabetta Mereu
Cellular Systems Genomics lab
Josep Carreras Leukaemia Research Institute, Badalona, Spain

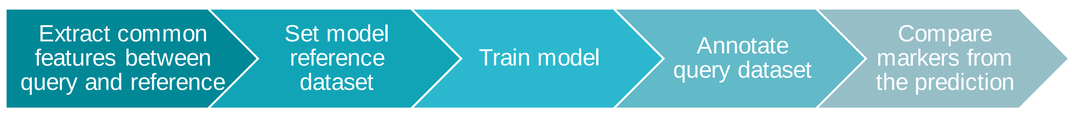
1/introduction

The validity of the populations identified with a clustering algorithm can only be determined by successful annotation of the encoded biology. Thanks to the recent curation and publication of large single-cell datasets in consortia such as the Human Cell Atlas, the biological annotation of a query dataset can be done via automatic label transfer from references [1].

deepSCore

/Fig1. DeepScore logo.

Deep learning models have demonstrated unparalleled performance and scalability in the annotation of large scale single cell datasets [2]. However, limitations such as the need of large computational resources hinders the practical application of these tools. Furthermore, these models are usually specific to a certain platform and modality. Here we present DeepSCore, a multi-language and multi-modality deep learning model for single-cell label transference. DeepSCore has been designed to be fast, scalable and user-friendly, with compatibility in both Seurat V4 [3] in R and AnnData [4] in Python.



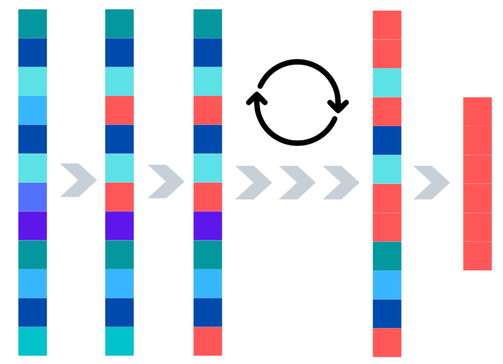
2/feature_selection

Features of input data must be shared among datasets, so finding a common set of features is a crucial first step:

- **Common genes** for scRNA and gene activites inferred from scATAC.
- **Common peaks** for scATAC.

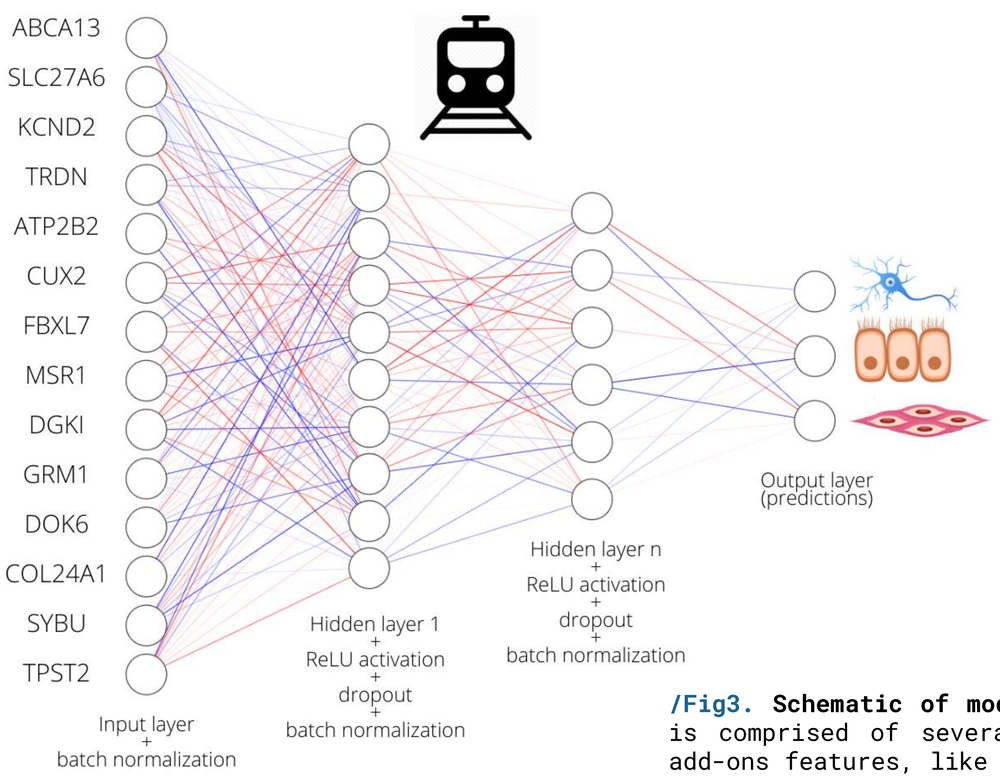
DeepSCore find a common set of features in an iterative manner:

- Find most variable genes or top peaks shared among datasets.
- Find reference markers in most variable genes or top peaks from the query dataset.
- Find reference markers in the query features.



/Fig2. Iterative feature selection. The algorithm will iteratively look for features to match the desired number specified by the user.

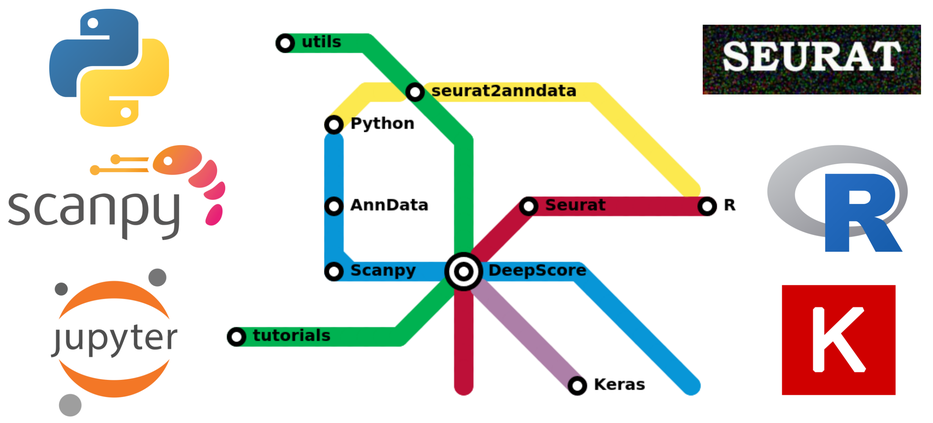
3/the_model



Add-on features

- Batch normalization
- Dropout
- Weight normalization
- L1 and L2 regularizers
- Early Stopping
- Learning Rate scheduler
- Tensorboard
- Export training as GIF

/Fig3. Schematic of model architecture. The model layout is comprised of several dense layers with customizable add-ons features, like batch normalization or dropout.



/Fig4. Multi-language metro map of the main different platforms and packages that DeepScore integrates.

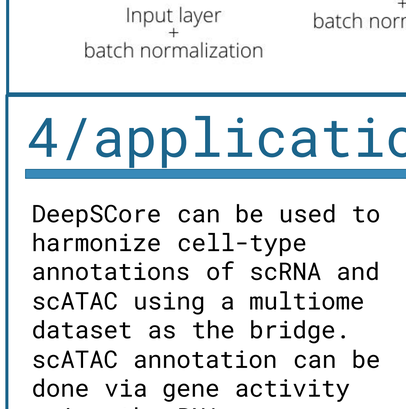
```
ds <- DeepScore(hidden_nodes=c(512, 256, 64),
                common_features=common_genes,
                n_labels=12) %>%
      set_reference(reference=ref_rna,
                    labels=ref_rna$anno) %>%
      train()
my_data <- annotate(ds, query=my_data)
```

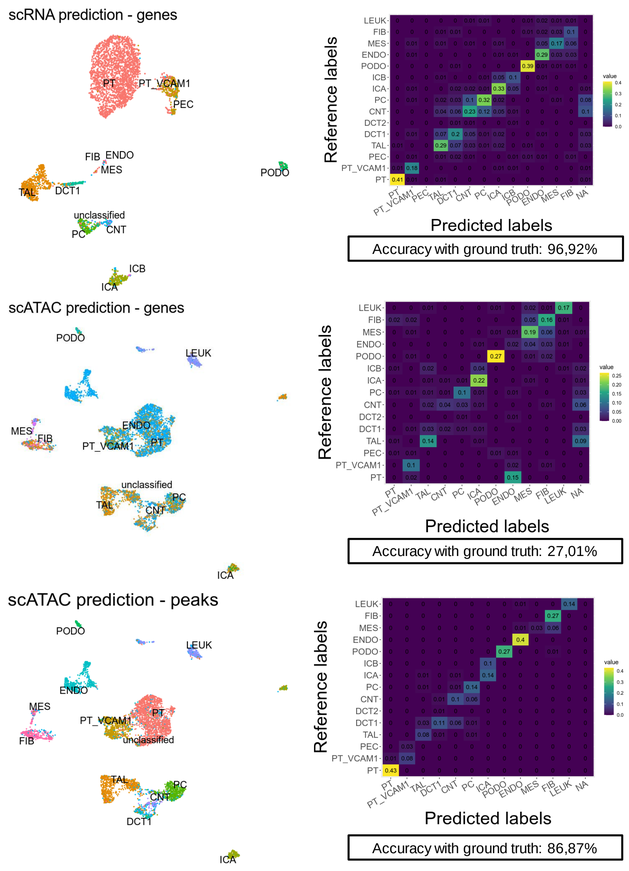
/Fig5. Example of DeepSCore implementation. Instantiate, train, and use your DeepSCore model in only 7 neat lines of code.

4/application_on_multiomic_data

DeepSCore can be used to harmonize cell-type annotations of scRNA and scATAC using a multiome dataset as the bridge. scATAC annotation can be done via gene activity using the RNA assay or peaks directly.

Human kidney single-cell data was produced in the scope of Chan Zuckerberg Initiative consortium, inside HCA.

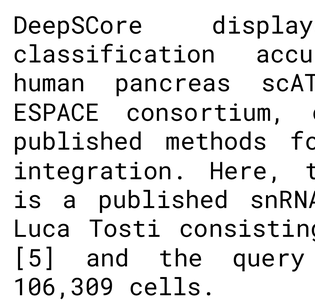


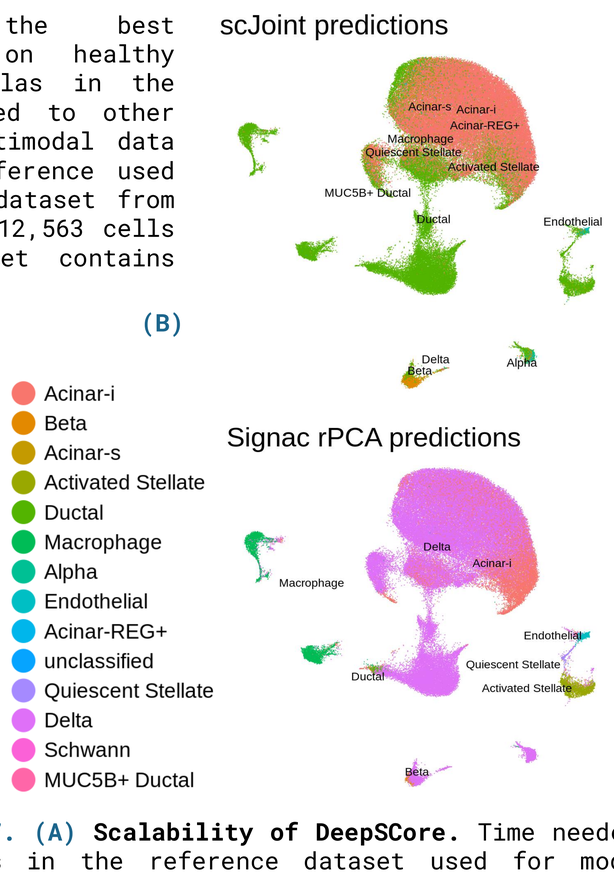


/Fig7. (A) Scalability of DeepSCore. Time needed and accuracy per number of cells in the reference dataset used for model training. The model was trained with: epochs=1, hidden layers=2, batch_size=32, dropout=50%.
(B) Comparison of deepSCore with published methods for label transfer.

5/application_on_scATAC_atlas

DeepSCore displayed the best classification accuracy on healthy human pancreas scATAC atlas in the ESPACE consortium, compared to other published methods for multimodal data integration. Here, the reference used is a published snRNA-seq dataset from Luca Tosti consisting of 112,563 cells [5] and the query dataset contains 106,309 cells.





/Fig7. (A) Scalability of DeepSCore. Time needed and accuracy per number of cells in the reference dataset used for model training. The model was trained with: epochs=1, hidden layers=2, batch_size=32, dropout=50%.
(B) Comparison of deepSCore with published methods for label transfer.

CHAN ZUCKERBERG INITIATIVE

/Fig7. Annotation of single assay scRNA-seq and scATAC-seq datasets using 10X multiome data as the bridge.

Institut de Recerca CONTRA LA LEUCÈMIA Josep Carreras

HUMAN CELL ATLAS

espace THE HCA PANCREAS INITIATIVE

/references

[1] Malte Luecken and Fabian Theis, 2019. *"Current best practices in single-cell RNA-seq analysis: a tutorial."*

[2] Giovanni Pasquini et al., 2021. *"Automated methods for cell type annotation on scRNA-seq data."*

[3] Yuhan Hao et al., 2021. *"Integrated analysis of multimodal single-cell data."*

[4] Isaac Virshup et al., 2021. *"Anndata: Annotated Data."*

[5] Luca Tosti et al., 2021. *"Single-nucleus and In Situ RNA-sequencing reveal cell topographies in the Human Pancreas."*

/future_directions

In the roadmap for DeepSCore we have in mind:

- Include support for other modalities such as spatial transcriptomics.
- Add functionalities to facilitate working with scATAC-seq datasets.
- Implement proposals from the community.

/acknowledgements

We thank Aina Rill and Ane Martinez for helpful advice and suggestions on the poster, as well as members of the Cellular Systems Genomics lab for insightful discussion. We also thank CZI and ESPACE consortia for data availability.

TRACK THE LIVE LEARNING OF A DEEPSCORE MODEL ON SINGLE CELL KIDNEY DATA HERE!

github.com/pabloswfly
linkedin.com/pabswfly
@pabloswfly