



DOI: [10.55458/neurolibre.00010](https://doi.org/10.55458/neurolibre.00010)



Reproducible Preprint

- [Jupyter Book](#) 

Code

- [Technical Screening](#) 
- [Submitted Repository](#) 

Reproducibility Assets

- [Repository](#) 
- [Dataset](#) 
- [Jupyter Book](#) 
- [Container](#) 

Moderator: [Agah Karakuzu](#) 

Screener(s):

- [@agahkarakuzu](#)

Submitted: 15 July 2022

Published: 22 October 2023

License

Authors of papers retain copyright  
and release the work under a  
Creative Commons Attribution 4.0  
International License ([CC BY 4.0](#)).

# Parcellating the parcellation issue - a proof of concept for reproducible analyses using Neurolibre

Pierre Bellec <sup>1,2</sup>, Saâd Jbabdi <sup>3</sup>, and R. Cameron Craddock <sup>4</sup>

1 Université de Montréal, Montréal, Canada 2 Centre de recherche de l'université de Montréal, Montréal, CA 3 University of Oxford, Oxford, UK 4 brainhack.org



THIS PDF IS INTENDED FOR CONTENT REGISTRATION PURPOSES ONLY! FOR FULL ACCESS AND INTERACTIVE READING OF THIS PUBLICATION, PLEASE VISIT [THE REPRODUCIBLE PREPRINT](#).

## Summary

Back in 2017, a special issue on the topic of **brain parcellation and segmentation** was published in the journal *Neuroimage*. We acted as guest editors for this special issue, and wrote an editorial ([Craddock et al., 2018](#)) providing an overview of all papers, sorted into categories. The categories were generated using a data-driven parcellation analysis, based on the words contained in the abstract of the articles. This jupyter book will allow interested readers to reproduce this analysis, as a proof of concept for reproducible publications using [jupyter books](#) and the [Neurolibre](#) preprint server.

## Acknowledgements

NeuroLibre is sponsored by the Canadian Open Neuroscience Platform (CONP), Brain Canada, Cancer Computers, the Courtois foundation, the Quebec Bioimaging Network, and Healthy Brains for Healthy Life.



## NOTE

**NOTE:** The following section in this document repeats the narrative content exactly as found in the [corresponding NeuroLibre Reproducible Preprint \(NRP\)](#). The content was automatically incorporated into this PDF using the NeuroLibre publication workflow ([Karakuzu et al., 2022](#)) to credit the referenced resources. The submitting author of the preprint has verified and approved the inclusion of this section through a GitHub pull request made to the [source repository](#) from which this document was built. Please note that the figures and tables have been excluded from this (static) document. **To interactively explore such outputs and re-generate them, please visit the corresponding NRP.** For more information on integrated research objects (e.g., NRPs) that bundle narrative and executable content for reproducible and transparent publications, please refer to DuPre et al. (2022). NeuroLibre is sponsored by the Canadian Open Neuroscience Platform (CONP) ([Harding et al., 2023](#)).

## Text mining

### List of papers

We first assembled the title, the name of the corresponding author, and the abstract for all the articles into a tabular-separated values (tsv) file, which we publicly archived on [Figshare](#). We use the [Repo2Data](#) tool developed by the NeuroLibre team to collect these data and include them in our reproducible computational environment.

### Word features

For each paper, we used [scikit-learn](#) ([Kramer, 2016](#)) to extract a bag of words representation for each abstract, picking on the 300 most important terms seen across all articles based on a term frequency-inverse document frequency ([tf-idf](#)) index. Following that, a special value decomposition was used to further reduce the dimensionality of the abstracts to 10 components. We ended up with a component matrix of dimension 38 (articles) times 10 (abstract text components). The distribution of each of the 38 articles across the 10 components is represented below. Note how some articles have particular high loadings on specific components, suggesting these may capture particular topics. Rather than visually inspect the component loadings to group paper ourselves, we are going to resort to an automated parcellation (clustering) technique.

### Parcellate the papers

Now that the content of each paper has been condensed into only 10 (hopefully informative) numbers, we can run these features into a trusted, classic parcellation algorithm: Ward's agglomerative hierarchical clustering, as implemented in the [scipy](#) library. We cut the hierarchy to extract 7 "paper parcels", and also use the hierarchy to re-order the papers, such that similar papers are close in order, as illustrated in a dendrogram representation.

### Similarity matrix

So, to get a better feel of the similarity between papers that was fed into the clustering procedure, we extracted the 38x38 (papers x papers) correlation matrix across features. Papers are re-ordered in the matrix according to the above hierarchy. Each "paper parcel" has been indicated by a white square along the diagonal, which represents the similarity measures between papers falling into the same parcel.

### Word cloud

Now, each paper of the special issue has been assigned to one and only one out of 7 possible "paper parcel". For each paper parcel, we can evaluate which words contribute more to the dominant component associated with that parcel.

### Categories

Thanks to the word clouds, these simple data-driven categories turned out to be fairly easily interpretable. For example, the word cloud of the category number 4 features prominently words like "white", "matter" and "bundles". If we examine the exact list of papers included in this category, we see that it is composed of four papers, which all considered parcels derived from white matter bundles with diffusion imaging. We can also check the distribution of component loadings for this category alone. As expected, there is a certain similarity in the component loadings for these papers, in particular along component 4:

## References

- Craddock, R. C., Bellec, P., & Jbabdi, S. (2018). Neuroimage special issue on brain segmentation and parcellation - editorial. *Neuroimage*, 170, 1–4. <https://doi.org/10.1016/j.neuroimage.2017.11.063>
- DuPre, E., Holdgraf, C., Karakuzu, A., Tetrel, L., Bellec, P., Stikov, N., & Poline, J.-B. (2022). Beyond advertising: New infrastructures for publishing integrated research objects. *PLOS Computational Biology*, 18(1), e1009651. <https://doi.org/10.1371/journal.pcbi.1009651>
- Harding, R. J., Bermudez, P., Bernier, A., Beauvais, M., Bellec, P., Hill, S., Karakuzu, A., Knoppers, B. M., Pavlidis, P., Poline, J.-B., Roskams, J., Stikov, N., Stone, J., Strother, S., Consortium, C., & Evans, A. C. (2023). The Canadian Open Neuroscience Platform—An open science framework for the neuroscience community. *PLOS Computational Biology*, 19(7), 1–14. <https://doi.org/10.1371/journal.pcbi.1011230>
- Karakuzu, A., DuPre, E., Tetrel, L., Bermudez, P., Boudreau, M., Chin, M., Poline, J.-B., Das, S., Bellec, P., & Stikov, N. (2022). *NeuroLibre : A preprint server for full-fledged reproducible neuroscience*. OSF Preprints. <https://doi.org/10.31219/osf.io/h89js>
- Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies* (pp. 45–53). Springer International Publishing. [https://doi.org/10.1007/978-3-319-33383-0\\_5](https://doi.org/10.1007/978-3-319-33383-0_5)