github.com/con/opfvta-reexecution

Neuroimaging Article Reexecution and Reproduction Assesment System

Horea-Ioan Ioanas[1], Austin Macdonald[1], Yaroslav O. Halchenko[1]

[1]Center for Open Neuroscience, Department of Psychological and Brain Sciences, Dartmouth College

Abstract

The value of research articles is increasingly contingent on data analysis results which substantiate their claims. Unlike data production steps, data analysis steps lend themselves to a higher standard of both transparency and repeated operator-independent execution. This higher standard can be approached via fully reexecutable research outputs, which contain the entire instruction set for end-to-end generation of an entire article solely from the earliest feasible provenance point, in a programatically executable format. In this study, we make use of a peer-reviewed neuroimaging article which provides complete but fragile reexecution instructions, as a starting point to formulate a new reexecution system which is both robust and portable. We render this system modular as a core design aspect, so that reexecutable article code, data, and environment specifications could potentially be substituted or adapted. In conjunction with this system, which forms the demonstrative product of this study, we detail the core challenges with full article reexecution and specify a number of best practices which permitted us to mitigate them. We further show how the capabilities of our system can subsequently be used to provide reproducibility assessments, both via simple statistical metrics and by visually highlighting divergent elements for human inspection. We argue that fully reexecutable articles are thus a feasible best practice, the usage of which can greatly enhance the understanding of data analysis variability and thus reliability of results. Lastly, we comment at length on the outlook for reexecutable research outputs and encourage re-use and derivation of the system produced herein.

Workflow

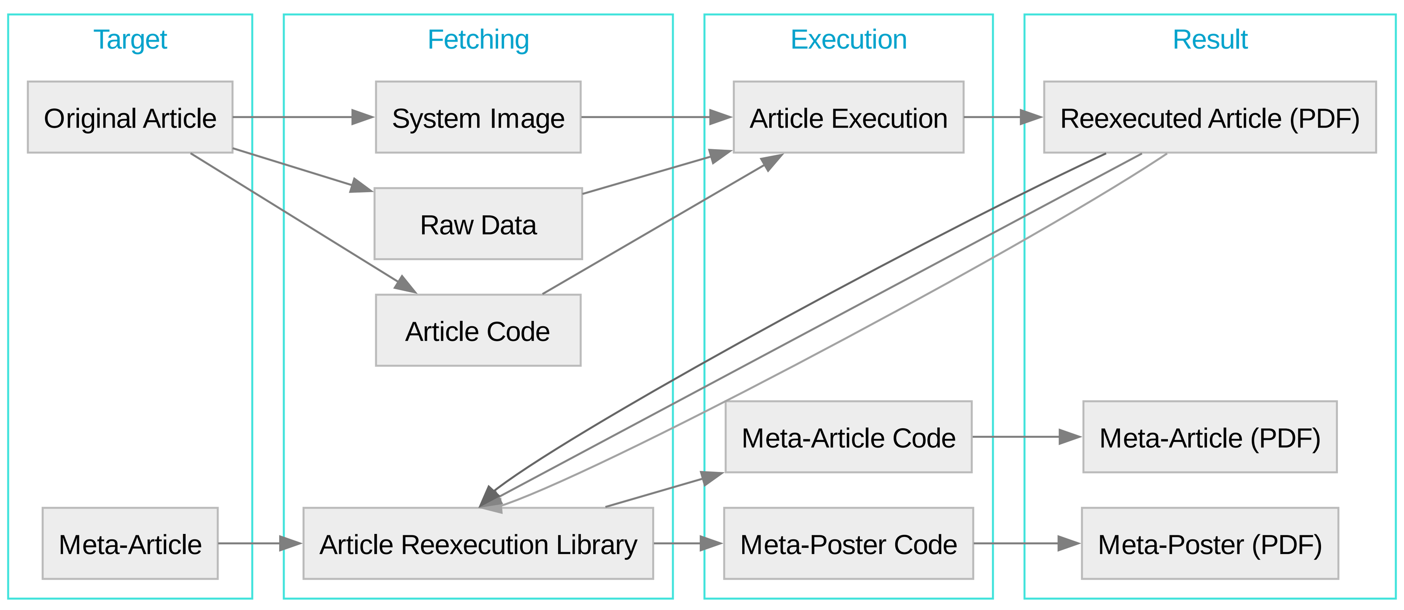


Figure 1: The reexecution system encompasses both the original article (first target), and the "meta-article" publishing materials (article manuscript, as well as this poster), the latter of which takes user- and developer-submitted reexecution results as an input for the reproduction quality assessment.

Topoplogy

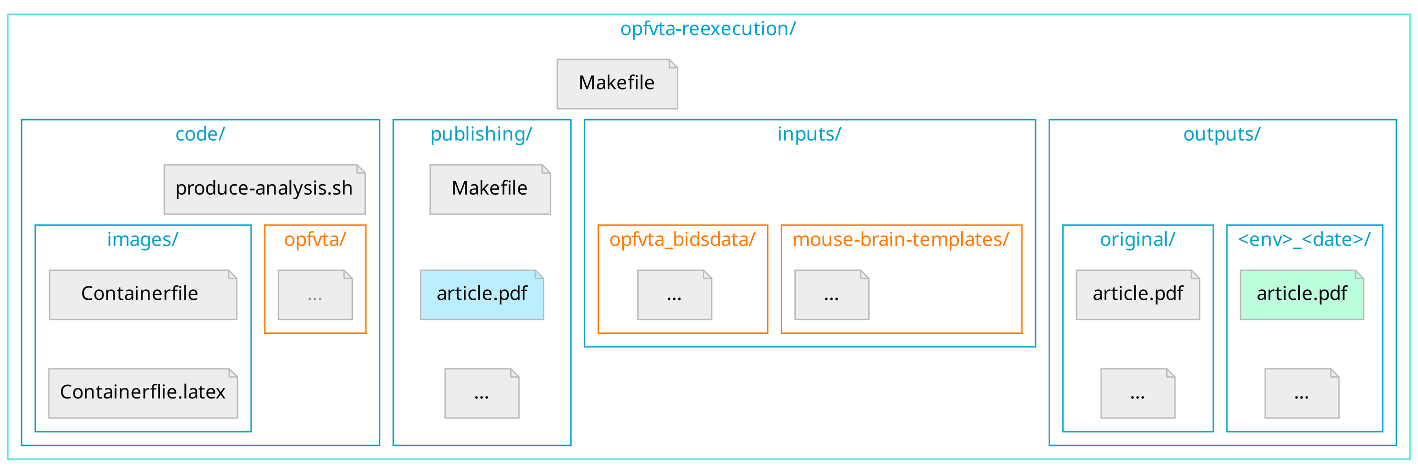


Figure 2: The reexecution workflow is supported by a resource topology in which reexecution code (first box), "meta-article" code (second box), reexecution resources (third box), and the reexecution output record (last box) are separated at the top directory level. The figure depicts directory trees via nested boxes, with external resources automatically fetched as via the reexecution code being highlighted in orange. The green highlighted article represents a sample reexecution output, and the blue highlighted article represents the manuscript, an analogous output to this poster generated in the same directory.

Best Practice Guidelines

As part of setting up an encompassing reexecution system, we formulate a number of best practices, including:

- **Errors should be fatal more often than not.**
  `set -eu`, prepended to POSIX shell scripts, will ensure that workflows fail when a subcommand does, or when an encountered variable is undefined.
- **Avoid assuming a directory context for execution.**
  `cd "$(dirname "$0")"`, prepended to POSIX shell scripts, will ensure that in complex workflows scripts can operate relative to their location directory context and not the execution context.
- **Workflow granularity greatly benefits efficiency.**
  While the underlying execution system of the target article, RepSeP [1] separates data analysis into two distinct (voxel-space "low iteration" and top-level "high iteration") steps, further granularity can benefit debugging, particularly in container environments.
- **Resources should be bundled into a DataLad superdataset.**
  Resource bundling, with usage of submodules for external resources (as seen in fig. 2) allows management of required resources via Git and associated technologies, such as DataLad [2] — this is known as the YODA principle [3].
- **Dependency versions inside container environments should be frozen as soon as feasible.**
  This is best accomplished via a package manager which uses version tracking for its software provision index; in Gentoo Linux, used here on account of broad provision of neuroscience packages [4], this can be done via:
  `cd /.../myrepo; git fetch origin $myhash; git checkout $myhash`.

References

[1] H.-I. Ioanas and M. Rudin, "Reproducible self-publishing for Python-based research," EuroSciPy, Aug. 2018.

[2] Y. Halchenko, K. Meyer, B. Poldrack, D. Solanky, A. Wagner, J. Gors, D. MacFarlane, D. Pustina, V. Sochat, S. Ghosh, C. Mönch, C. Markiewicz, L. Waite, I. Shlyakhter, A. de la Vega, S. Hayashi, C. Häusler, J.-B. Poline, T. Kadelka, K. Skytén, D. Jarecka, D. Kennedy, T. Strauss, M. Cieslak, P. Vavra, H.-I. Ioanas, R. Schneider, M. Pflüger, J. Haxby, S. Eickhoff, and M. Hanke, *DataLad: distributed system for joint management of code, data, and their relationship*, vol. 6. The Open Journal, July 2021.

[3] M. Hanke, M. Visconti di Oleggio Castello, K. Meyer, B. Poldrack, and Y. O. Halchenko, "YODA: YODA's organigram on data analysis." Poster presented at the annual meeting of the Organization for Human Brain Mapping, Singapore, 2018.

[4] H.-I. Ioanas, B. Saab, and M. Rudin, "Gentoo linux for neuroscience — a replicable, flexible, scalable, rolling-release environment that provides direct access to development software," *Research Ideas and Outcomes*, vol. 3, p. e12095, Feb. 2017.

Reproduction Assessment Showcase

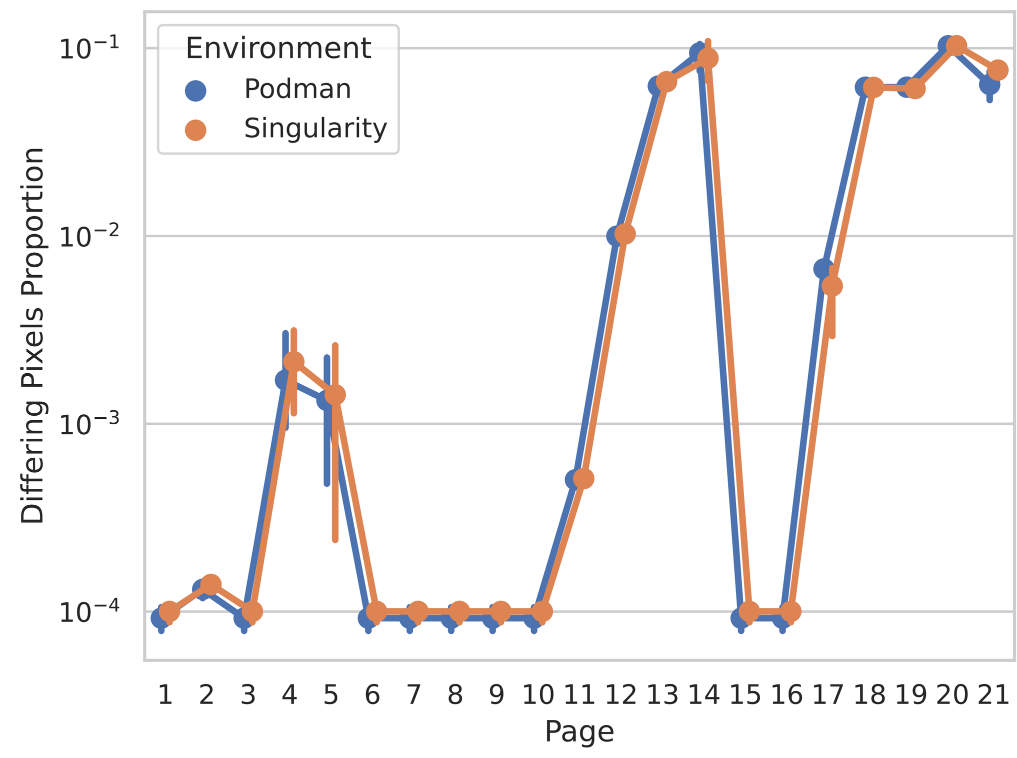


Figure 3: Page-wise voxel difference comparison across multiple reexecutions in different environments indicates consistency of variability in both extent and location.

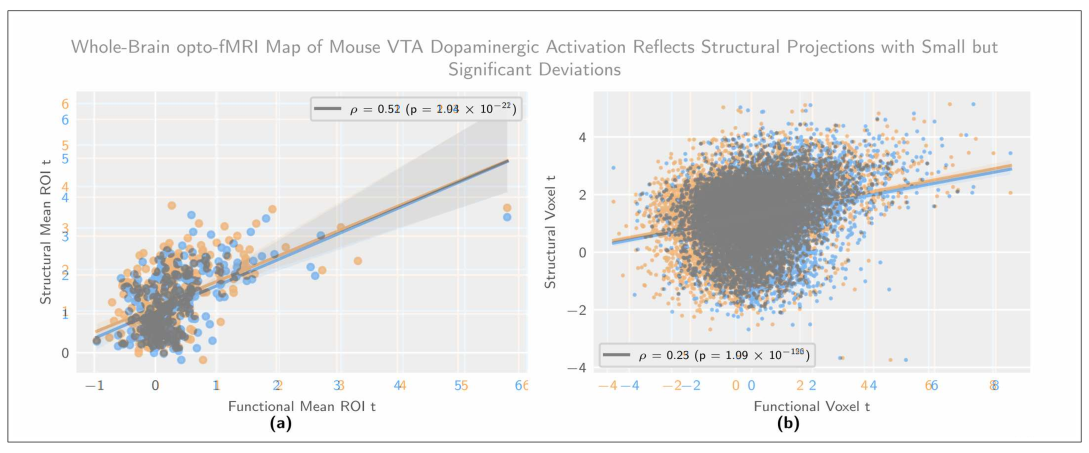
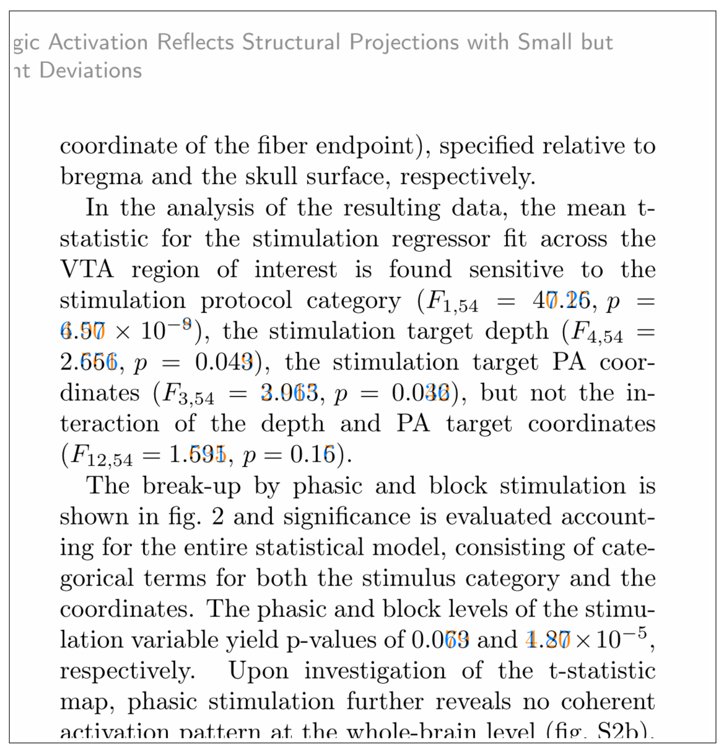


Figure 4: One notable source of variability are data plots, where it can be observed that even as data points vary to an almost full extent, statistical summaries can remain constant.



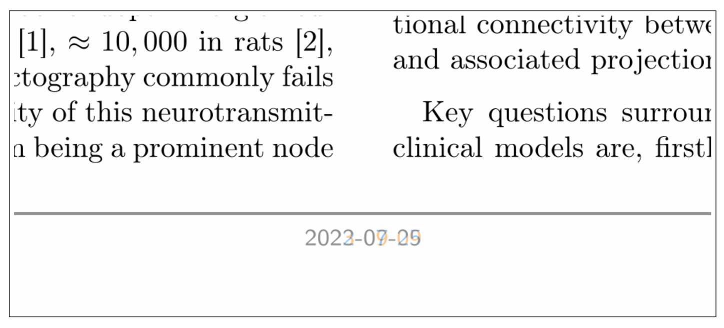
gic Activation Reflects Structural Projections with Small but
nt Deviations

coordinate of the fiber endpoint), specified relative to bregma and the skull surface, respectively.

In the analysis of the resulting data, the mean t-statistic for the stimulation regressor fit across the VTA region of interest is found sensitive to the stimulation protocol category ($F_{1,54} = 47.26$, $p = 6.57 \times 10^{-8}$), the stimulation target depth ($F_{4,54} = 2.656$, $p = 0.043$), the stimulation target PA coordinates ($F_{3,54} = 3.063$, $p = 0.036$), but not the interaction of the depth and PA target coordinates ($F_{12,54} = 1.595$, $p = 0.16$).

The break-up by phasic and block stimulation is shown in fig. 2 and significance is evaluated accounting for the entire statistical model, consisting of categorical terms for both the stimulus category and the coordinates. The phasic and block levels of the stimulation variable yield p-values of 0.063 and 4.87×10^{-5}, respectively. Upon investigation of the t-statistic map, phasic stimulation further reveals no coherent activation pattern at the whole-brain level (fig. S2b).

Figure 5: Text differences in statistical summaries account for a small proportion of voxel differences, but can remain well-localized instead of spreading via test shift if statistical summaries are appropriately trimmed down to a constant length.



[1], ≈ 10,000 in rats [2],
ctography commonly fails
ity of this neurotransmit-
n being a prominent node

tional connectivity betw
and associated projectio

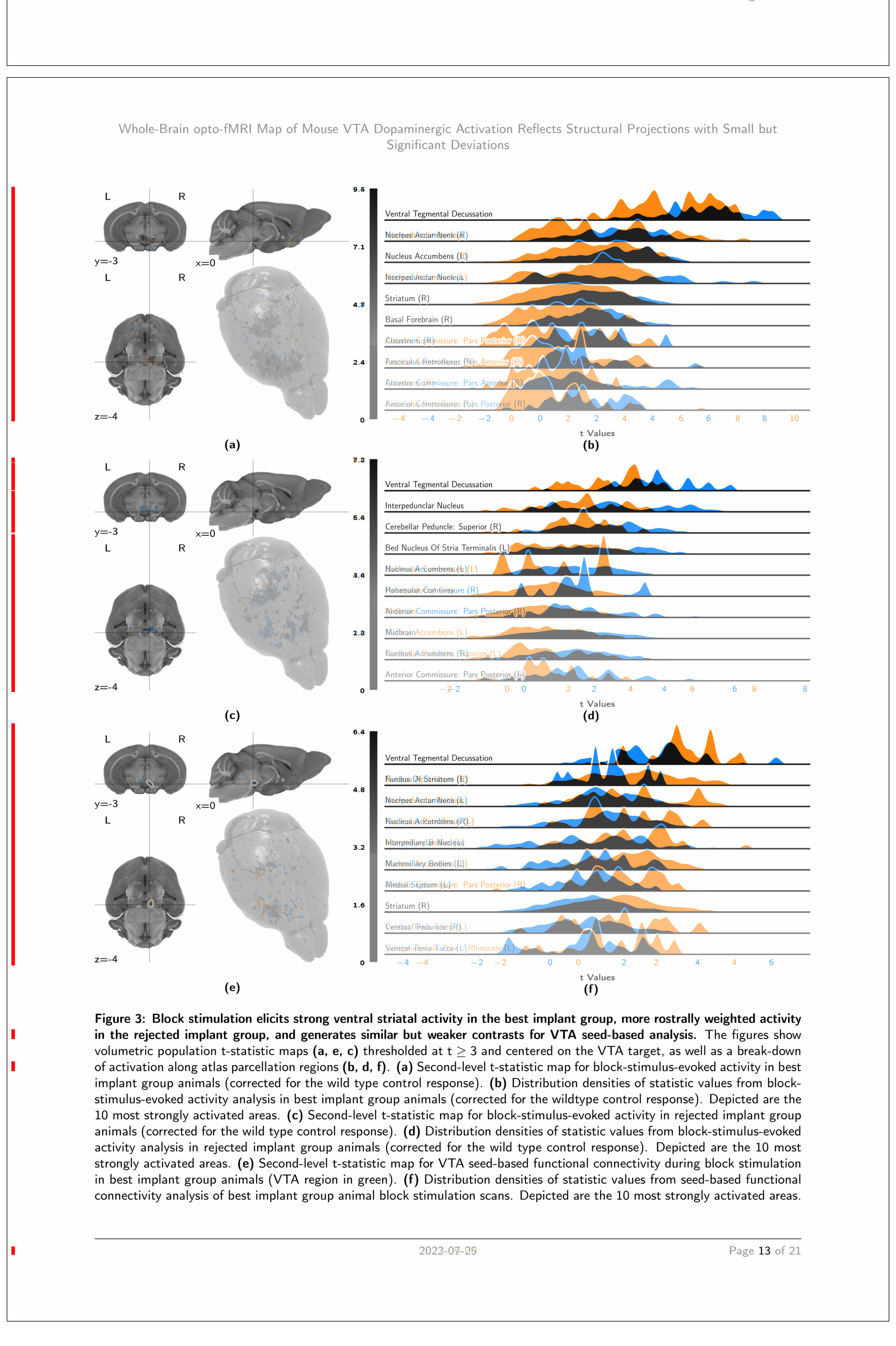
Key questions surrou
clinical models are, firstl

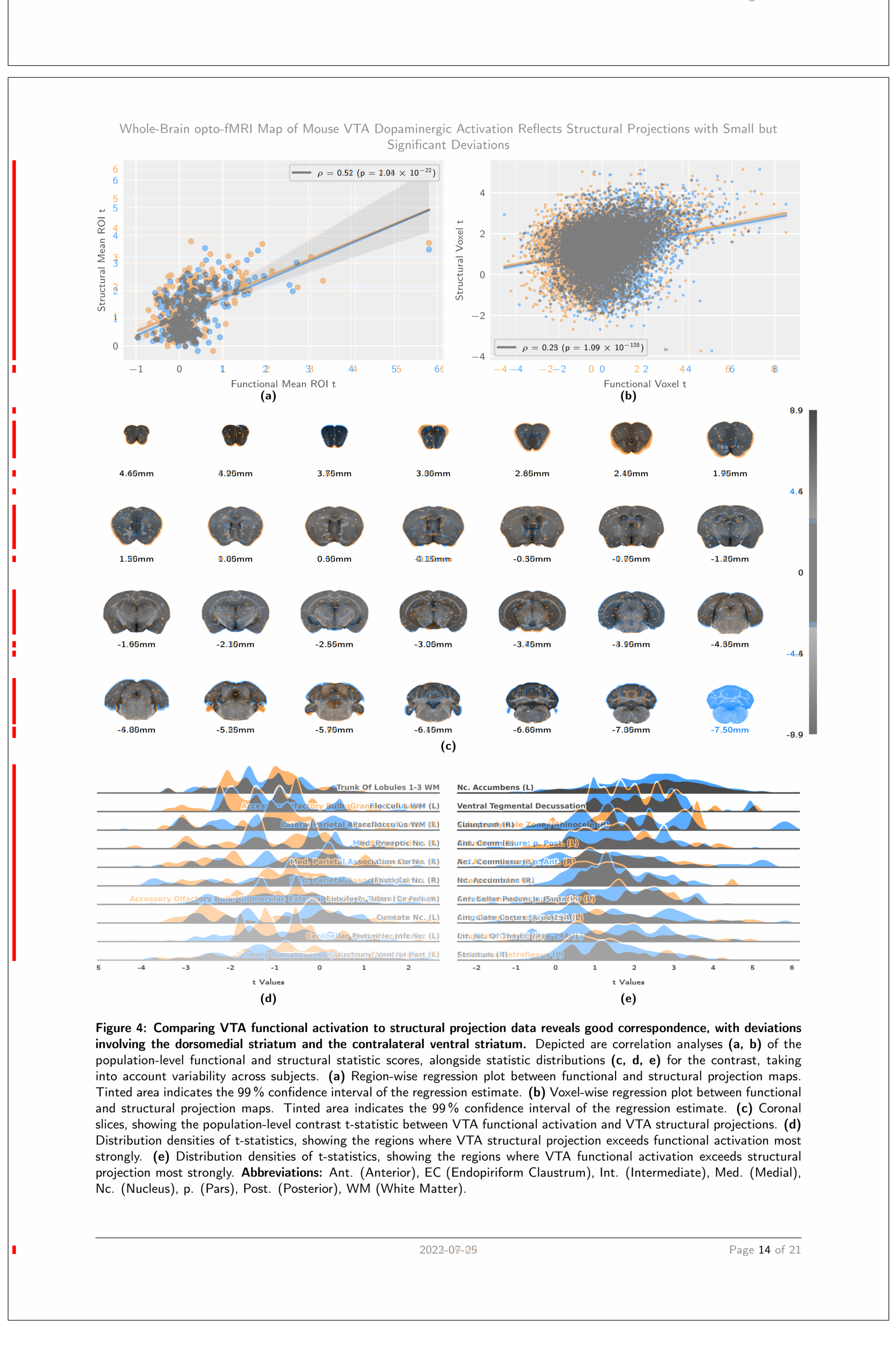
2023-07-25

Figure 6: A good litmus test for monitoring differences (accounting for the baseline difference in fig. 3) is the datestamp of the reexecution, which should always be expected to differ from the manuscript.

Full Document Comparison

Reproduction assessment is based on *full* document "diffs". The following figures are excerpts, with tinted highlighting (blue for the original manuscript, and orange for reexecution result). First row pages exemplify inline statistical differences and second row pages exemplify figure differences. Differing sections are highlighted with a red left-hand marking.





DOI: 10.5281/zenodo.3477675

Society for Neuroscience — 2023-11-10

uni@chymera.eu