The Journal of Open Source Software

# Interactive exploration of benchmarking results with *bettr*

**Charlotte Soneson** [1,2¶], **Federico Marini** [3,4], **Daniel Incicau** [2,5], **Anthony Sonrel** [2,5], **Almut Lütge** [6], **Reto Gerber** [2,5], **Ben Carrillo** [2,5], **Izaskun Mallona** [2,5], **and Mark D Robinson** [2,5]

**1** Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland **2** SIB Swiss Institute of Bioinformatics, Basel, Switzerland **3** Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany **4** Research Center for Immunotherapy (FZI) Mainz, Mainz, Germany **5** Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland **6** Swiss Data Science Centre, Zurich, Switzerland ¶ Corresponding author

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:** Chris Vernon

**Reviewers:**

- @srvanderplas

**Submitted:** 12 December 2025
**Published:** unpublished



## Summary

Method benchmarking is a core part of many research fields and aims to establish best practices in method selection and application as well as help identifying gaps and possibilities for improvement in existing methods. A typical benchmarking study scores a set of methods using a variety of different metrics, intended to capture different aspects of performance and usability. For practical purposes, e.g., when a user needs to select a method for their own analysis, it is often the case that not all of the evaluated metrics are equally important. For example, the ability of a statistical method to handle arbitrarily complex experimental designs may not be relevant if the user only needs to perform a two-group comparison, while memory frugality may be of high importance if the analysis will be performed on a system with limited computational resources. Hence, having the ability to easily and adaptively create a weighted summary score by which the methods can be ranked, or alternatively to create a visual summary of only the desired metrics, would make it possible for users to tailor their method choice based on the aspects that are most relevant to them.

Inspired by the OECD 'Better Life Index', which allows users to score OECD countries by a customizable weighted average of various factors contributing to human well-being, we developed the *bettr* R/Bioconductor package to provide similar support for consumers of benchmarking studies. Given a table with values of evaluation metrics across the evaluated methods, as well as optional method annotations and metric groupings, *bettr* allows users to visualize performance summaries emphasizing the aspects and evaluation metrics that are most important to them (Figure 1). Examples of metrics include the area under the receiver operating characteristic (ROC) curve for a classification task, or the execution time. Method annotations (e.g., whether they are designed to work on untransformed or log-transformed input data) allows the user to evaluate whether methods with certain characteristics tend to perform better than others. *bettr* can be used interactively as an R/Shiny application (Chang et al., 2025), or programmatically by calling the underlying functions directly for full reproducibility and integration into analytical pipelines (see the package vignettes for more details).

The Journal of Open Source Software

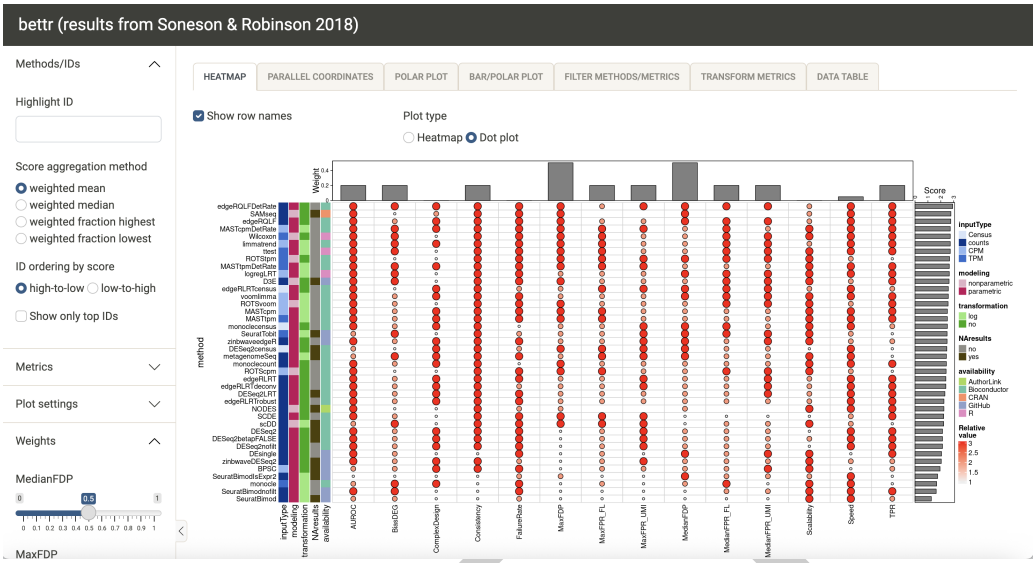


**Figure 1:** Screenshot of a *bettr* application illustrating the benchmarking results from Soneson & Robinson (2018). The size and color of the circles indicate the evaluation score for each method (rows) and metric (columns), with higher values being more favorable. The bars above the plot indicate the current weights for the metrics (these are controlled by the sliders in the bottom of the left sidebar), and the bars to the right of the plot represent the aggregated scores for the methods (in this case, the weighted mean across metrics). The colored bars to the left of the plot represent categorical annotations and characterizations of the methods, which are not taken into account when determining the ranking. In addition to the displayed dot plot, the results can be visualized in several other ways, including a heatmap, polar plots, and using parallel coordinates. Moreover, the set of methods and metrics to include in the display can be controlled by the user via the 'Filter methods/metrics' tab.

*bettr* accepts input in multiple formats, including a collection of data frames, a SummarizedExperiment object (Morgan et al., 2025) or a JSON file, and contains functions for converting one input format to another. This increases the flexibility of the application and makes it easy to combine with a variety of workflows. For example, a user working locally in R may find it more convenient to generate a set of data frames, while an automated benchmarking workflow could export all necessary components in a single, platform-independent JSON file. *bettr* can also be deployed in server mode, allowing users to upload their own input data (in JSON format) to a running app.

## Statement of Need

The interactive nature of *bettr* makes it particularly suitable for visual, exploratory analysis of benchmarking results, since the user can interactively modify the weights associated with the different evaluation metrics and immediately see how it affects the ranking of the methods. It is also straightforward for authors of benchmarking studies to deploy an instance of *bettr* using, e.g., a local Shiny server or a commercial option such as https://shinyapps.io, to allow readers to easily explore their results.

## State of the Field

Complementary functionality for creating summary representations (e.g., heatmap-like visualizations) of benchmarking results is provided e.g. by the funkyheatmap package (Cannoodt et al., 2025), and benchmarking platforms such as OpenEBench (Capella-Gutierrez et al., 2017) and OpenProblems (Luecken et al., 2025) also produce result visualizations

for the included benchmarks. However, existing tools typically lack flexibility in either input format or means of deployment, or the ability to explore results interactively, and are not intended to support user-specific metric weighting. To the best of our knowledge, *bettr* is the first generic tool for benchmark visualization that combines interactivity and flexibility in metric weighting with ease of use and a transparent, programmatic interface.

## Software Design

The design philosophy behind *bettr* focuses on flexibility, accessibility, and reproducibility. By supporting several input formats, from collections of data frames via a single R-based object to an all-encompassing JSON file, *bettr* can be used for downstream analysis and visualization of results generated using a wide range of benchmark setups and programming languages. *bettr* is fully open source and easily installable as an R package, mainly distributed via Bioconductor with the most recent development version also available on GitHub. In addition, it is available via *r-universe*, which among other things provides binaries for several platforms, including WebAssembly. As a complement to the interactive interface, full reproducibility is enabled by the equivalent programmatic interface to the functionality, as well as the ability to export the processed data as either a shareable csv file containing the metric values as well as the final score, or an R list that can be used directly as the input for further analysis and visualization.

## Research Impact Statement

In 2025, *bettr* was downloaded almost 3,000 times (by 1,593 unique IPs) from Bioconductor alone. Adoption of *bettr* is expected to increase further in the near future, as it has become the first supported integration for automated metrics reporting within the Omnibenchmark project. Omnibenchmark (Mallona et al., 2026) is a benchmarking system that automates and standardizes routine aspects of benchmarking through standardization and formalization of benchmarking plans. During execution, Omnibenchmark collects computational performance metrics (e.g., peak memory usage, CPU utilization, run time, etc), as well as, where applicable, algorithmic performance metrics (e.g., F1 scores, ARIs, etc). These results are then exported via a command-line interface in a JSON format compatible with *bettr*, enabling automated reporting. This integration lowers the barrier for benchmark authors to produce rich, interactive summaries of complex benchmarking studies without requiring custom visualization pipelines. Hence, broader adoption of Omnibenchmark as a benchmarking framework is expected to further drive the use of *bettr* as an interactive platform for exploring and interpreting benchmark results.

## Availability and Examples

*bettr* is available via Bioconductor and on GitHub. A collection of example data sets and *bettr* configurations are available from https://github.com/csoneson/bettr-examples. In addition, an example instance, using data from Soneson & Robinson (2018), is deployed on https://csoneson.shinyapps.io/soneson2018de/.

## AI Usage Disclosure

The majority of the *bettr* code base was developed over several years, in a public GitHub repository, without the use of generative AI. For one pull request (#25), adding the capabilities to support JSON files as input and to cache the state of the app, Claude Sonnet 4.5 (Anthropic) was used to assist with drafting implementation of new functions and UI components based on predefined design documents, while maintaining existing functionality as is. At least two of the package developers reviewed the code carefully, edited it where necessary, and verified that the

new code performed as intended and did not introduce regressions in existing functionality (which was further verified using the comprehensive set of existing unit tests).

## Acknowledgements

CS is supported by the Novartis Research Foundation. The work of FM is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 318346496. MDR acknowledges funding from the Swiss National Science Foundation (grants 200021_212940 and 310030_204869) as well as support from swissuniversities P5 Phase B funding (project 23-36_14). The funders did not have any role in the design of the study, the collection, analysis and interpretation of data, or in writing the manuscript.

## References

Cannoodt, R., Deconinck, L., Couckuyt, A., Markov, N. S., Zappia, L., Luecken, M. D., Interlandi, M., Saeys, Y., & Saelens, W. (2025). Funkyheatmap: Visualising data frames with mixed data types. *Journal of Open Source Software*, *10*(108), 7698. https://doi.org/10.21105/joss.07698

Capella-Gutierrez, S., Iglesia, D. de la, Haas, J., Lourenco, A., Fernández, J. M., Repchevsky, D., Dessimoz, C., Schwede, T., Notredame, C., Gelpi, J. L., & Valencia, A. (2017). Lessons learned: Recommendations for establishing critical periodic scientific benchmarking. *bioRxiv*. https://doi.org/10.1101/181677

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Aden-Buie, G., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2025). *shiny: Web Application Framework for R*. https://doi.org/10.32614/CRAN.package.shiny

Luecken, M. D., Gigante, S., Burkhardt, D. B., Cannoodt, R., Strobl, D. C., Markov, N. S., Zappia, L., Palla, G., Lewis, W., Dimitrov, D., Vinyard, M. E., Magruder, D. S., Mueller, M. F., Andersson, A., Dann, E., Qin, Q., Otto, D. J., Klein, M., Botvinnik, O. B., … Krishnaswamy, S. (2025). Defining and benchmarking open problems in single-cell analysis. *Nature Biotechnology*, *43*(7), 1035–1040. https://doi.org/10.1038/s41587-025-02694-w

Mallona, I., Luetge, A., Carrillo, B., Incicau, D., Gerber, R., Meara, A., Sonrel, A., Soneson, C., & Robinson, M. D. (2026). Omnibenchmark: Transparent, reproducible, extensible and standardized orchestration of solo and collaborative benchmarks. *arXiv [q-Bio.OT]*. https://doi.org/10.48550/arXiv.2409.17038

Morgan, M., Obenchain, V., Hester, J., & Pagès, H. (2025). *SummarizedExperiment: A container (S4 class) for matrix-like assays*. https://doi.org/10.18129/B9.bioc.SummarizedExperiment

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, *15*(4), 255–261. https://doi.org/10.1038/nmeth.4612