SimOmics: A Simulation Toolkit for Multivariate and Multi-Omics Data

Kaitao Lai[1]

8 July 2025

[1] University of Sydney

1 Summary

SimOmics is an R package designed to generate realistic, multivariate, and multi-omics synthetic datasets. It is intended for use in benchmarking, method development, and reproducibility in bioinformatics, particularly in the context of omics integration tasks such as those encountered in transcriptomics, proteomics, and metabolomics. SimOmics supports latent factor simulation, sparsity structures, block-wise covariance modeling, and biologically inspired noise models and feature dimensions.

2 Statement of need

Simulated datasets play a crucial role in statistical method development, especially in fields such as systems biology and multi-omics integration, where complex relationships often exist between multiple data modalities. However, existing simulation tools are either too simple (e.g., based on basic multivariate distributions) or lack support for biologically relevant structures such as sparsity, block integration, or shared latent variables.

SimOmics fills this gap by providing an accessible yet powerful framework to generate synthetic multi-omics datasets that resemble real biological complexity. This is critical for the testing and benchmarking of methods like mixOmics (Rohart et al. 2017), MOFA2 (Argelaguet et al. 2020), and iCluster (Wang et al. 2014).

3 Features

- Simulate multiple omics blocks with flexible dimensions
- Inject shared latent factors or independent noise
- Control inter-block correlation via a block covariance structure
- Add Gaussian noise and customize signal-to-noise ratio
- Plot PCA, correlation, and integration results
- Export data and integrate with other packages like `mixOmics`

4 Why Simulated Data Matters

Simulated datasets—when designed to reflect realistic biological complexity—are essential tools in bioinformatics research (Zhang et al. 2020; Huang et al. 2017). SimOmics enables users to generate such datasets for a range of multivariate and multi-omics scenarios.

4.1 Advantages of synthetic datasets:

- **Method development**: Early-stage algorithm development benefits from controllable data without needing access to real datasets (Witten, Tibshirani, and Hastie 2009).
- **Known ground truth**: Simulated data allow researchers to benchmark models based on sensitivity, specificity, and overall performance with known signal.
- **Stress-testing**: Users can design edge-case scenarios (e.g., high noise, low correlation, small sample size) not typically found in public datasets (Teschendorff and Relton 2018).
- **Reproducibility**: Synthetic data ensures that benchmarking can be reproduced across institutions and software implementations.

5 Use cases and target audience

SimOmics is intended for:

- Researchers developing new multi-omics integration methods
- Developers of machine learning models for high-dimensional data
- Authors writing benchmarking papers that require known ground truth
- Bioinformatics instructors demonstrating omics integration techniques

Users include PhD students, statisticians, and software developers in computational biology. Although it is not aimed at experimental biologists, it is a valuable tool for method validation and reproducible research.

6 Related work

Several existing R packages such as `clusterGeneration`, `bnlearn`, and `simuPOP` support general-purpose simulation of data. However, none offer block-wise omics simulation with latent factors and realistic biological noise models. SimOmics builds on the need observed in published tools such as mixOmics (Rohart et al. 2017), MOFA2 (Argelaguet et al. 2020), and iCluster (Wang et al. 2014) to benchmark model performance in structured settings.

7 Example Use Case

The following figure shows the result of applying `mixOmics::block.plsda()` to two synthetic omics blocks (transcriptome and proteome) generated by SimOmics. Class labels A and B are partially separable, reflecting shared latent structure while preserving variability.

Figure 1. PLS-DA of the simulated multi-omics dataset. Samples (colored by class A and B) are projected into a shared latent space. The partial overlap illustrates integrated but not fully class-discriminative structure — a common challenge in real-world omics data.

8 Software Repository

The source code for SimOmics is freely available on GitHub at:
https://github.com/biosciences/SimOmics

9 Acknowledgements

This project was initiated to support method development and reproducible benchmarking in multi-omics research. We acknowledge the contributions of the broader open-source and mixOmics communities.

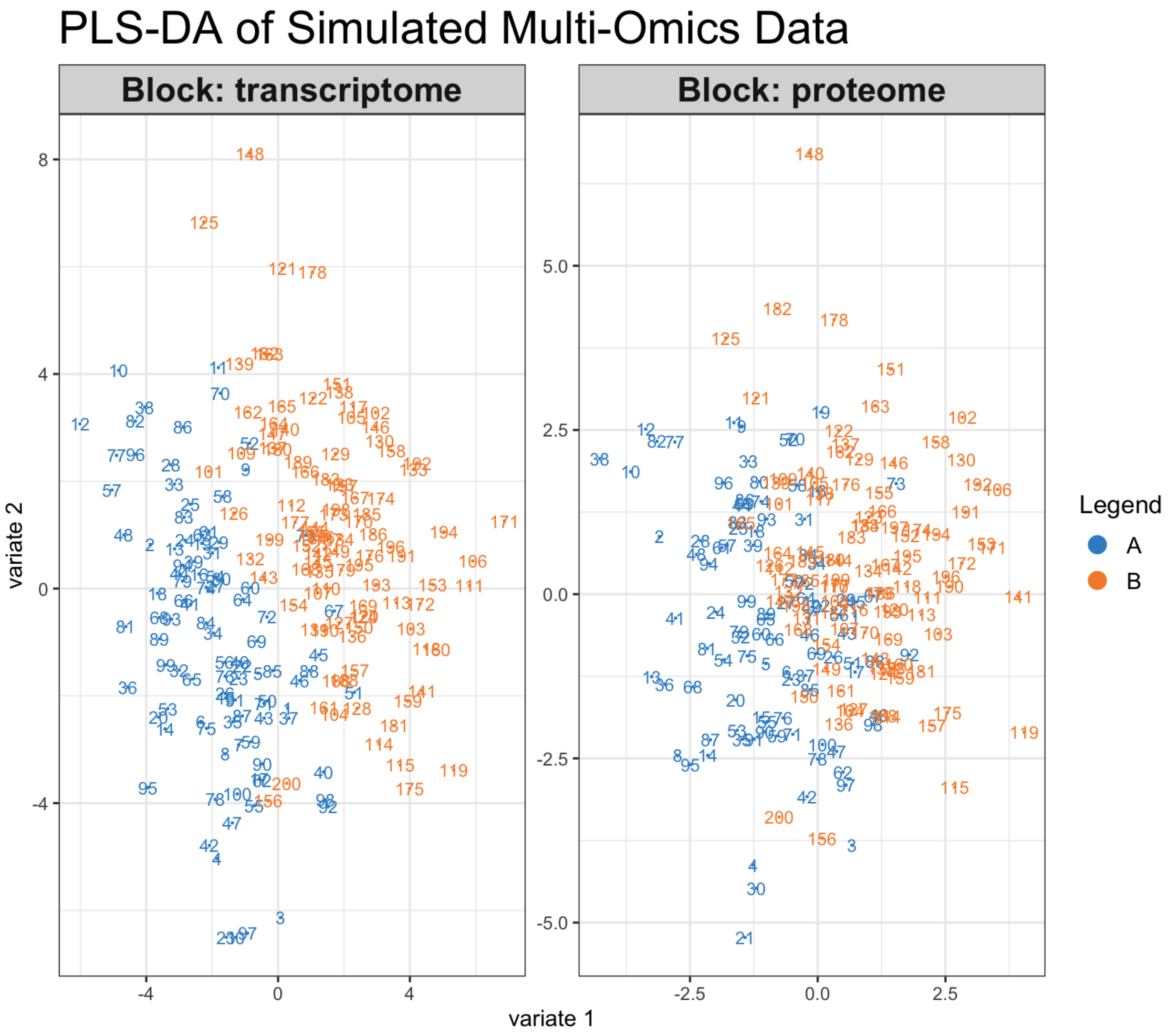

Figure 1: PLS-DA of simulated multi-omics dataset

References

Argelaguet, Ricard, Denis Arnol, Dmitry Bredikhin, Yann Deloro, Lars Velten, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2020. "MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data." *Genome Biology* 21: 111. https://doi.org/10.1186/s13059-020-02015-1.

Huang, Jing, Guoliang Zhou, Shujie Meng, Yan Chen, Tingting Dong, Zhi Cheng, and Yu Wang. 2017. "Systematic Evaluation of Molecular Networks for Discovery of Disease Genes." *Cell Systems* 5 (5): 460–70. https://doi.org/10.1016/j.cels.2017.09.006.

Rohart, Florian, Benoit Gautier, Amrit Singh, and Kim-Anh Lê Cao. 2017. "mixOmics: An r Package for 'Omics Feature Selection and Multiple Data Integration." *PLoS Computational Biology* 13 (11): e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Teschendorff, Andrew E, and Caroline L Relton. 2018. "Statistical and Integrative System-Level Analysis of DNA Methylation Data." *Nature Reviews Genetics* 19 (3): 129–47. https://doi.org/10.1038/nrg.2017.86.

Wang, Bo, Asif M Mezlini, Fahad Demir, Marc Fiume, Zhaleh Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. 2014. "iClusterPlus: Integrative Clustering of Multiple Genomic Data Types with Feature Selection." *Statistical Applications in Genetics and Molecular Biology* 13 (6): 511–26. https://doi.org/10.1515/sagmb-2013-0059.

Witten, Daniela M, Robert Tibshirani, and Trevor Hastie. 2009. "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis." *Biostatistics* 10 (3): 515–34. https://doi.org/10.1093/biostatistics/kxp008.

Zhang, Honghan, Yi Wang, Qi Wang, Wenjie Zheng, Zhaonan Huang, and Zhen Huang. 2020. "Synthetic Data Generation and Its Application in Machine Learning for Healthcare." *Npj Digital Medicine* 3: 106. https://doi.org/10.1038/s41746-020-00343-7.