The Journal of Open Source Software

SpatialProteomicsNet: A unified interface for spatial proteomics data access for computer vision and machine learning

Adriano Martinelli [1,2,4] and **Marianna Rapsomaniki** [1,3,4]

1 University Hospital Lausanne (CHUV), Lausanne, Switzerland **2** ETH Zurich, Zurich, Switzerland **3** University of Lausanne (UNIL), Lausanne, Switzerland **4** Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

`SpatialProteomicsNet` is an open-source Python package that provides a harmonized and standardized interface for accessing spatial proteomics and multiplexed imaging datasets, including imaging mass cytometry (IMC) (Giesen et al., 2014) and multiplexed ion beam imaging time-of-flight (MIBI-TOF) (Keren et al., 2019) data. The package enables researchers to load raw spatially-resolved proteomics data from multiple studies in a unified format, apply and retrieve data structures ready for downstream machine learning analysis or model training. By focusing on open-source raw data processing and enforcing common data schemas (e.g., standardized image and single-cell data formats), `SpatialProteomicsNet` promotes reproducible and efficient research in computational and spatial biology. The library is designed to serve the broader community working on spatial proteomics by easing data access and integration into machine learning workflows.

Statement of Need

Spatially-resolved proteomics, recently named Nature Method of the Year 2024 ("Method of the Year 2024," 2024), enable the quantification of proteins in single cells within their tissue context, revealing intricate aspects of spatial cellular arrangement and communication. In the context of cancer, these advancements provide unprecedented insights into the heterogeneity of the tumor and its microenvironment, and the underlying mechanisms affecting tumor initiation, progression, and response to treatment (Lewis et al., 2021). IMC and MIBI-TOF are among the most popular technologies, with dozens of high-dimensional datasets made publicly available per year. The increasing availability of these datasets has fueled algorithmic development in machine learning and computer vision. Numerous models that perform a variety of tasks, such as cell segmentation (Greenwald et al., 2022), cell type annotation (Geuenich et al., 2021), representation learning (Wenckstern et al., 2025) or heterogeneity analysis (Martinelli & Rapsomaniki, 2022) tailored to spatial proteomics data have been recently developed, with corresponding widely used packages.

However, a critical gap hindering model development, reproducibility and cross-study analyses is the lack of unified frameworks to access and process the data. Spatial proteomics datasets, often deposited in public repositories such as Zenodo(European Organization For Nuclear Research & OpenAIRE, 2013) or Figshare (*Figshare - Credit for All Your Research*, n.d.), typically contain a collection of components, such as raw and preprocessed images, segmentation masks, extracted single-cell intensities, panel descriptions and associated clinical metadata, uploaded in disparate, non-standardized formats (e.g., mixed .tiff, .csv, custom JSONs), with varying metadata structures and inconsistent preprocessing that vary greatly between studies and labs.

Working with these fragmented datasets implies a significant time investment for researchers to locate and download the data, and write custom scripts to handle their specific data structure, creating barriers to entry, complicating usage and hindering robust benchmarking. While existing data frameworks developed by the spatial transcriptomics community such as SpatialData (Marconato et al., 2025) and Pysodb (Yuan et al., 2023) are gaining popularity and can be extended to spatial proteomics, they often come with heavier dependencies and general-purpose abstractions that may be unnecessarily complex for researchers focused on fast, standardized access to real-world IMC or MIBI-TOF datasets.

`SpatialProteomicsNet` is an open-source Python package that addresses these gaps by:

- Providing a lightweight, unified interface to widely-used curated spatial proteomics datasets.
- Abstracting dataset-specific structure, letting users access data components (images, masks, metadata) through a consistent schema.
- Supporting reproducible preprocessing via modular, reusable interfaces for common pipeline steps.
- Facilitating integration in machine learning and computer vision models by streamlining dataset loading into standard formats.
- Encouraging community contributions for expanding and maintaining harmonized dataset access.

This unified approach allows scientists to abstract away dataset-specific idiosyncrasies and focus on biological and analytical questions rather than data wrangling. `SpatialProteomicsNet` is intentionally minimal, tailored to machine learning and computer vision workflows (e.g., loading images, masks, and cell-level metadata with minimal setup) without depending on larger ecosystem packages (e.g., anndata, xarray, zarr, dask). `SpatialProteomicsNet` gives immediate access to curated datasets with ready-to-use utilities, eliminating the need to write custom loaders or parse inconsistent formats. As such, it is particularly friendly to the growing community of ML developers, researchers, and engineers entering the emerging field of spatial biology. By harmonizing data access, our package enables more straightforward integration of spatial proteomics data into machine learning and modeling frameworks, ultimately accelerating biomedical discovery.

Supported Datasets

The package supports the following public spatial proteomics datasets:

- **Keren et al. 2018** – MIBI-TOF of triple-negative breast cancer (Keren et al., 2018)
- **Jackson et al. 2020** – IMC of breast cancer (Jackson et al., 2020)
- **Danenberg et al. 2022** – IMC of breast cancer (Danenberg et al., 2022)
- **Cords et al. 2024** – IMC of NSCLC (Cords et al., 2024)

name	images	masks	markers	annotated cells	clinical samples
Danenberg2022	794	794	39	1123466	794
Cords2024	2070	2070	43	5984454	2072
Jackson2020	735	735	35	1224411	735
Keren2018	41	41	36	201656	41

Table 1: Summary statistics of supported spatial proteomics datasets in the package.

Each dataset is accessible through a standardized class interface that mimics the pytorch lightning (Falcon & team, 2019) philosophy and includes methods for downloading, preparing, and accessing processed components (images, masks, features and metadata). These datasets follow consistent naming conventions and data schemas, making them immediately usable for downstream tasks.

Conclusion

`SpatialProteomicsNet` lowers the technical barrier to working with spatial proteomics data by providing unified, open access to several published datasets and processing routines. Its modular design and standardized outputs make it a practical tool for researchers developing computational methods in spatial biology. We welcome contributions and extensions from the community and envision this package as a foundation for reproducible spatial proteomics analysis.

Acknowledgements

We thank Prof. Raza Ali, Prof. Leeat Keren, Prof. Michael Angelo and Dr. Lena Cords for providing detailed information and facilitating access to the corresponding datasets. This project has been made possible in part by grant number 2024-345909 from the Chan-Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

References

Cords, L., Engler, S., Haberecker, M., Rüschoff, J. H., Moch, H., De Souza, N., & Bodenmiller, B. (2024). Cancer-associated fibroblast phenotypes are associated with patient outcome in non-small cell lung cancer. *Cancer Cell*, *42*(3), 396–412.e5. https://doi.org/10.1016/j.ccell.2023.12.021

Danenberg, E., Bardwell, H., Zanotelli, V. R. T., Provenzano, E., Chin, S. F., Rueda, O. M., Green, A., Rakha, E., Aparicio, S., Ellis, I. O., Bodenmiller, B., Caldas, C., & Ali, H. R. (2022). Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature Genetics*, *54*(5), 660–669. https://doi.org/10.1038/s41588-022-01041-y

European Organization For Nuclear Research, & OpenAIRE. (2013). *Zenodo*. CERN. https://doi.org/10.25495/7GXK-RD71

Falcon, W., & team, T. P. L. (2019). *PyTorch lightning* (Version 1.4). https://doi.org/10.5281/zenodo.3828935

Figshare - credit for all your research. (n.d.). https://figshare.com/.

Geuenich, M. J., Hou, J., Lee, S., Ayub, S., Jackson, H. W., & Campbell, K. R. (2021). Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data. *Cell Systems*, *12*(12), 1173–1186.e5. https://doi.org/10.1016/j.cels.2021.08.012

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D., & Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, *11*(4), 417–422. https://doi.org/10.1038/nmeth.2869

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C. C., McIntosh, B. J., Leow, K. X., Schwartz, M. S., Pavelchek, C., Cui, S., Camplisson, I., Bar-Tal, O., Singh, J., Fong, M., Chaudhry, G., Abraham, Z., Moseley, J., … Van Valen, D. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, *40*(4), 555–565. https://doi.org/10.1038/s41587-021-01094-0

Jackson, H. W., Fischer, J. R., Zanotelli, V. R. T., Ali, H. R., Mechera, R., Soysal, S. D., Moch, H., Muenst, S., Varga, Z., Weber, W. P., & Bodenmiller, B. (2020). The single-cell pathology landscape of breast cancer. *Nature*, *578*(7796), 615–620. https:

//doi.org/10.1038/s41586-019-1876-x

Keren, L., Bosse, M., Thompson, S., Risom, T., Vijayaragavan, K., McCaffrey, E., Marquez, D., Angoshtari, R., Greenwald, N. F., Fienberg, H., Wang, J., Kambham, N., Kirkwood, D., Nolan, G., Montine, T. J., Galli, S. J., West, R., Bendall, S. C., & Angelo, M. (2019). MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science Advances*, *5*(10), eaax5851. https://doi.org/10.1126/sciadv.aax5851

Keren, L., Marquez, D., Bosse, M., Angoshtari, R., Jain, S., Varma, S., Yang, S. R., Kurian, A., Van Valen, D., West, R., Bendall, S. C., & Angelo, M. (2018). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, *174*(6), 1373–1387.e19. https://doi.org/10.1016/j.cell.2018.08.039

Lewis, S. M., Asselin-Labat, M.-L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., Merino, D., Rogers, K. L., & Naik, S. H. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods*, *18*(9), 997–1012. https://doi.org/10.1038/s41592-021-01203-6

Marconato, L., Palla, G., Yamauchi, K. A., Virshup, I., Heidari, E., Treis, T., Vierdag, W.-M., Toth, M., Stockhaus, S., Shrestha, R. B., Rombaut, B., Pollaris, L., Lehner, L., Vöhringer, H., Kats, I., Saeys, Y., Saka, S. K., Huber, W., Gerstung, M., … Stegle, O. (2025). SpatialData: An open and universal data framework for spatial omics. *Nature Methods*, *22*(1), 58–62. https://doi.org/10.1038/s41592-024-02212-x

Martinelli, A. L., & Rapsomaniki, M. A. (2022). ATHENA: Analysis of tumor heterogeneity from spatial omics measurements. *Bioinformatics*, *38*(11), 3151–3153. https://doi.org/10.1093/bioinformatics/btac303

Method of the Year 2024: Spatial proteomics. (2024). *Nature Methods*, *21*(12), 2195–2196. https://doi.org/10.1038/s41592-024-02565-3

Wenckstern, J., Jain, E., Vasilev, K., Pariset, M., Wicki, A., Gut, G., & Bunne, C. (2025). *AI-powered virtual tissues from spatial proteomics for clinical diagnostics and biomedical discovery* (No. arXiv:2501.06039). arXiv. https://doi.org/10.48550/arXiv.2501.06039

Yuan, Z., Pan, W., Zhao, X., Zhao, F., Xu, Z., Li, X., Zhao, Y., Zhang, M. Q., & Yao, J. (2023). SODB facilitates comprehensive exploration of spatial omics data. *Nature Methods*, *20*(3), 387–399. https://doi.org/10.1038/s41592-023-01773-7