

<sup>1</sup> **SpatialProteomicsNet**: A unified interface for spatial  
<sup>2</sup> proteomics data access for computer vision and  
<sup>3</sup> machine learning

<sup>4</sup> **Adriano Martinelli**  <sup>1,2,4</sup> and **Marianna Rapsomaniki**  <sup>1,3,4</sup>

<sup>5</sup> 1 University Hospital Lausanne (CHUV), Lausanne, Switzerland 2 ETH Zurich, Zurich, Switzerland 3  
<sup>6</sup> University of Lausanne (UNIL), Lausanne, Switzerland 4 Swiss Institute of Bioinformatics (SIB),  
<sup>7</sup> Lausanne, Switzerland

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

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

## Summary

<sup>9</sup> SpatialProteomicsNet is an open-source Python package that provides a harmonized and  
<sup>10</sup> standardized interface for accessing spatial proteomics and multiplexed imaging datasets,  
<sup>11</sup> including imaging mass cytometry ( IMC) ([Giesen et al., 2014](#)) and multiplexed ion beam  
<sup>12</sup> imaging time-of-flight (MIBI-TOF) ([Keren et al., 2019](#)) data. The package enables researchers  
<sup>13</sup> to load raw spatially-resolved proteomics data from multiple studies in a unified format,  
<sup>14</sup> apply and retrieve data structures ready for downstream machine learning analysis or model  
<sup>15</sup> training. By focusing on open-source raw data processing and enforcing common data schemas  
<sup>16</sup> (e.g., standardized image and single-cell data formats), SpatialProteomicsNet promotes  
<sup>17</sup> reproducible and efficient research in computational and spatial biology. The library is designed  
<sup>18</sup> to serve the broader community working on spatial proteomics by easing data access and  
<sup>19</sup> integration into machine learning workflows.

## Statement of Need

<sup>21</sup> Spatially-resolved proteomics, recently named Nature Method of the Year 2024 ("Method of  
<sup>22</sup> the Year 2024," [2024](#)), enable the quantification of proteins in single cells within their tissue  
<sup>23</sup> context, revealing intricate aspects of spatial cellular arrangement and communication. In the  
<sup>24</sup> context of cancer, these advancements provide unprecedented insights into the heterogeneity of  
<sup>25</sup> the tumor and its microenvironment, and the underlying mechanisms affecting tumor initiation,  
<sup>26</sup> progression, and response to treatment ([Lewis et al., 2021](#)). IMC and MIBI-TOF are among the  
<sup>27</sup> most popular technologies, with dozens of high-dimensional datasets made publicly available  
<sup>28</sup> per year. The increasing availability of these datasets has fueled algorithmic development  
<sup>29</sup> in machine learning and computer vision. Numerous models that perform a variety of tasks,  
<sup>30</sup> such as cell segmentation ([Greenwald et al., 2022](#)), cell type annotation ([Geuenich et al.,  
31](#) [2021](#)), representation learning ([Wenckstern et al., 2025](#)) or heterogeneity analysis ([Martinelli  
32](#) & Rapsomaniki, [2022](#)) tailored to spatial proteomics data have been recently developed, with  
<sup>33</sup> corresponding widely used packages.

<sup>34</sup> However, a critical gap hindering model development, reproducibility and cross-study analyses is  
<sup>35</sup> the lack of unified frameworks to access and process the data. Spatial proteomics datasets, often  
<sup>36</sup> deposited in public repositories such as Zenodo([European Organization For Nuclear Research  
37](#) & OpenAIRE, [2013](#)) or Figshare ([Figshare - Credit for All Your Research](#), n.d.), typically  
<sup>38</sup> contain a collection of components, such as raw and preprocessed images, segmentation masks,  
<sup>39</sup> extracted single-cell intensities, panel descriptions and associated clinical metadata, uploaded  
<sup>40</sup> in disparate, non-standardized formats (e.g., mixed .tiff, .csv, custom JSONs), with varying  
<sup>41</sup> metadata structures and inconsistent preprocessing that vary greatly between studies and labs.

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

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

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

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

## 72 Supported Datasets

73 The package supports the following public spatial proteomics datasets:

- 74     ▪ [Keren et al. 2018](#) – MIBI-TOF of triple-negative breast cancer ([Keren et al., 2018](#))
- 75     ▪ [Jackson et al. 2020](#) – IMC of breast cancer ([Jackson et al., 2020](#))
- 76     ▪ [Danenberg et al. 2022](#) – IMC of breast cancer ([Danenberg et al., 2022](#))
- 77     ▪ [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

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

79 Additionally, dummy datasets are provided to mimic real data structure for development and  
 80 testing purposes.

81 Each dataset is accessible through a standardized class interface that mimics the pytorch  
 82 lightning ([Falcon & team, 2019](#)) philosophy and includes methods for downloading, preparing,

83 and accessing processed components (images, masks, features and metadata). These datasets  
84 follow consistent naming conventions and data schemas, making them immediately usable for  
85 downstream tasks.

## 86 Conclusion

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

## 93 Acknowledgements

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

## 98 References

- 99 Cords, L., Engler, S., Haberecker, M., Rüschoff, J. H., Moch, H., De Souza, N., & Bodenmiller,  
100 B. (2024). Cancer-associated fibroblast phenotypes are associated with patient outcome in  
101 non-small cell lung cancer. *Cancer Cell*, 42(3), 396–412.e5. <https://doi.org/10.1016/j.ccr.2023.12.021>
- 103 Danenberg, E., Bardwell, H., Zanotelli, V. R. T., Provenzano, E., Chin, S. F., Rueda, O.  
104 M., Green, A., Rakha, E., Aparicio, S., Ellis, I. O., Bodenmiller, B., Caldas, C., & Ali,  
105 H. R. (2022). Breast tumor microenvironment structures are associated with genomic  
106 features and clinical outcome. *Nature Genetics*, 54(5), 660–669. <https://doi.org/10.1038/s41588-022-01041-y>
- 108 European Organization For Nuclear Research, & OpenAIRE. (2013). Zenodo. CERN. <https://doi.org/10.25495/7GXK-RD71>
- 110 Falcon, W., & team, T. P. L. (2019). PyTorch lightning (Version 1.4). <https://doi.org/10.5281/zenodo.3828935>
- 112 Figshare - credit for all your research. (n.d.). <https://figshare.com/>.
- 113 Geuenich, M. J., Hou, J., Lee, S., Ayub, S., Jackson, H. W., & Campbell, K. R. (2021).  
114 Automated assignment of cell identity from single-cell multiplexed imaging and proteomic  
115 data. *Cell Systems*, 12(12), 1173–1186.e5. <https://doi.org/10.1016/j.cels.2021.08.012>
- 116 Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler,  
117 P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D.,  
118 & Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular  
119 resolution by mass cytometry. *Nature Methods*, 11(4), 417–422. <https://doi.org/10.1038/nmeth.2869>
- 121 Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C. C.,  
122 McIntosh, B. J., Leow, K. X., Schwartz, M. S., Pavelchek, C., Cui, S., Camplisson, I.,  
123 Bar-Tal, O., Singh, J., Fong, M., Chaudhry, G., Abraham, Z., Moseley, J., ... Van Valen,  
124 D. (2022). Whole-cell segmentation of tissue images with human-level performance using  
125 large-scale data annotation and deep learning. *Nature Biotechnology*, 40(4), 555–565.

- 126        <https://doi.org/10.1038/s41587-021-01094-0>
- 127        Jackson, H. W., Fischer, J. R., Zanotelli, V. R. T., Ali, H. R., Mechera, R., Soysal, S.  
128        D., Moch, H., Muenst, S., Varga, Z., Weber, W. P., & Bodenmiller, B. (2020). The  
129        single-cell pathology landscape of breast cancer. *Nature*, 578(7796), 615–620. <https://doi.org/10.1038/s41586-019-1876-x>
- 130        Keren, L., Bosse, M., Thompson, S., Risom, T., Vijayaragavan, K., McCaffrey, E., Marquez,  
131        D., Angoshtari, R., Greenwald, N. F., Fienberg, H., Wang, J., Kambham, N., Kirkwood,  
132        D., Nolan, G., Montine, T. J., Galli, S. J., West, R., Bendall, S. C., & Angelo, M. (2019).  
133        MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure.  
134        *Science Advances*, 5(10), eaax5851. <https://doi.org/10.1126/sciadv.aax5851>
- 135        Keren, L., Marquez, D., Bosse, M., Angoshtari, R., Jain, S., Varma, S., Yang, S. R., Kurian, A.,  
136        Van Valen, D., West, R., Bendall, S. C., & Angelo, M. (2018). A Structured Tumor-Immune  
137        Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam  
138        Imaging. *Cell*, 174(6), 1373–1387.e19. <https://doi.org/10.1016/j.cell.2018.08.039>
- 139        Lewis, S. M., Asselin-Labat, M.-L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C.,  
140        Merino, D., Rogers, K. L., & Naik, S. H. (2021). Spatial omics and multiplexed imaging  
141        to explore cancer biology. *Nature Methods*, 18(9), 997–1012. <https://doi.org/10.1038/s41592-021-01203-6>
- 142        Marconato, L., Palla, G., Yamauchi, K. A., Virshup, I., Heidari, E., Treis, T., Vierdag, W.-M.,  
143        Toth, M., Stockhaus, S., Shrestha, R. B., Rombaut, B., Pollaris, L., Lehner, L., Vöhringer,  
144        H., Kats, I., Saeys, Y., Saka, S. K., Huber, W., Gerstung, M., ... Stegle, O. (2025).  
145        SpatialData: An open and universal data framework for spatial omics. *Nature Methods*,  
146        22(1), 58–62. <https://doi.org/10.1038/s41592-024-02212-x>
- 147        Martinelli, A. L., & Rapsomaniki, M. A. (2022). ATHENA: Analysis of tumor heterogeneity  
148        from spatial omics measurements. *Bioinformatics*, 38(11), 3151–3153. <https://doi.org/10.1093/bioinformatics/btac303>
- 149        Method of the Year 2024: Spatial proteomics. (2024). *Nature Methods*, 21(12), 2195–2196.  
150        <https://doi.org/10.1038/s41592-024-02565-3>
- 151        Wenckstern, J., Jain, E., Vasilev, K., Pariset, M., Wicki, A., Gut, G., & Bunne, C. (2025).  
152        *AI-powered virtual tissues from spatial proteomics for clinical diagnostics and biomedical*  
153        *discovery* (No. arXiv:2501.06039). arXiv. <https://doi.org/10.48550/arXiv.2501.06039>
- 154        Yuan, Z., Pan, W., Zhao, X., Zhao, F., Xu, Z., Li, X., Zhao, Y., Zhang, M. Q., & Yao, J.  
155        (2023). SODB facilitates comprehensive exploration of spatial omics data. *Nature Methods*,  
156        20(3), 387–399. <https://doi.org/10.1038/s41592-023-01773-7>