

# <sup>1</sup> scpviz: A Python bioinformatics toolkit for Single-cell Proteomics-omics analysis

<sup>3</sup> Marion Pang  <sup>1¶</sup>, Baiyi Quan  <sup>2</sup>, Ting-Yu Wang  <sup>2</sup>, and Tsui-Fen Chou  <sup>1,2¶</sup>

<sup>5</sup> 1 Division of Biology and Biological Engineering, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125 2 Proteome Exploration Laboratory, Beckman Institute, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125 ¶ Corresponding author

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Published: unpublished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](#))<sup>3</sup>

## <sup>8</sup> Summary

<sup>9</sup> Proteomics seeks to characterize protein dynamics by measuring both protein abundance and <sup>10</sup> post-translational modifications (PTMs), such as phosphorylation, acetylation, and ubiquitination, which regulate protein activity, localization, and interactions. In bottom-up proteomics <sup>11</sup> workflows, proteins are enzymatically digested into peptides that are measured as spectra, <sup>12</sup> from which these peptide-spectrum matches (PSMs) are aggregated to infer protein-level <sup>13</sup> identifications and quantitative abundance estimates. Analyzing the two levels of data at <sup>14</sup> both the peptide level (short fragments observed directly) and the protein level (assembled <sup>15</sup> from peptide evidence) in tandem is crucial for translating raw measurements into biologically <sup>16</sup> interpretable results.

<sup>18</sup> Single-cell proteomics extends these approaches to resolve protein expression at the level of individual cells or microdissected tissue regions. Such data are typically sparse, with many <sup>19</sup> missing values, and are generated within complex experimental designs involving multiple <sup>20</sup> classes of samples (e.g., cell type, treatment, condition). These properties distinguish single-cell <sup>21</sup> proteomics from bulk experiments and create unique challenges in data processing, normalization, <sup>22</sup> and interpretation. The single-cell transcriptomics community has established a mature <sup>23</sup> ecosystem for managing similar challenges, exemplified by the scanpy package ([Wolf et al., 2018](#)) <sup>24</sup> and the broader scverse ecosystem. Building on these foundations, scpviz extends <sup>25</sup> the AnnData data structure to the domain of proteomics. scpviz is a Python package <sup>26</sup> that streamlines single-cell and spatial proteomics workflows, supporting a complete analysis <sup>27</sup> pipeline from raw peptide-level data to protein-level summaries and downstream interpretation <sup>28</sup> through differential expression, enrichment analysis, and network analysis. At its core, scpviz <sup>29</sup> centers the pAnnData class, an AnnData-affiliated data structure specialized for proteomics. <sup>30</sup> Together, these components make scpviz a comprehensive and extensible framework for <sup>31</sup> single-cell proteomics. By combining flexible data structures, reproducible workflows, and <sup>32</sup> seamless integration with the AnnData/Scanpy ecosystem, the package enables researchers to <sup>33</sup> efficiently connect peptide-level evidence to protein-level interpretation, thereby accelerating <sup>34</sup> methodological development and biological discovery in proteomics.

## <sup>36</sup> Statement of need

<sup>37</sup> Although general-purpose data analysis frameworks such as scanpy ([Wolf et al., 2018](#)) and <sup>38</sup> the broader scverse ecosystem have become indispensable for single-cell transcriptomics, <sup>39</sup> comparable tools for proteomics remain limited. Existing proteomics software often focus on <sup>40</sup> specialized tasks (e.g., peptide identification or spectrum assignment) and do not provide a <sup>41</sup> unified framework for downstream analysis of peptide- and protein-level data within single-cell

42 and spatial contexts. `scpviz` is designed to address these gaps by offering an integrated  
43 system for the complete proteomics workflow, from raw peptide-level evidence to protein-level  
44 summaries and biological interpretation. The package is intended for computational biologists  
45 and proteomics researchers working with low-input or single-cell datasets. The package is  
46 designed to support the complete analysis pipeline, extending from raw peptide-level data to  
47 protein-level summaries, and biological interpretation (e.g., differential expression, enrichment  
48 analysis, network analysis).

49 At the core of `scpviz` is the `pAnnData` class, an `AnnData`-affiliated data structure specialized  
50 for proteomics. It organizes peptide (`.pep`) and protein (`.prot`) `AnnData` objects together  
51 with supporting attributes such as `.summary`, `.metadata`, `.rs` matrices (protein-peptide re-  
52 lationships), and `.stats`. This design allows users to move flexibly between peptide- and  
53 protein-level perspectives, while preserving compatibility with established Python libraries for  
54 data science and visualization.

55 The package extends beyond simple data storage by implementing a wide array of proteomics-  
56 specific functions. These include filtering operations (e.g., requiring proteins to be supported  
57 by at least two unique peptides), normalization and imputation strategies tailored for sparse  
58 datasets, and visualization methods such as PCA, UMAP, clustermaps, and violin or box  
59 plots of abundance distributions. For downstream interpretation, `scpviz` integrates with  
60 external resources: UniProt for protein annotation and STRING for functional enrichment and  
61 protein-protein interaction network analysis. `scpviz` `pAnnData` objects also integrate seamlessly  
62 with the `scipy` package (Wolf et al., 2018); for example, the `scpviz.pAnnData.clean_x()`  
63 function prepares data matrices in the appropriate format for direct use in `Scipy` workflows.

64 The design philosophy of `scpviz` emphasizes both usability and extensibility. General users can  
65 rely on its streamlined API to import, process, and visualize single-cell proteomics data without  
66 deep programming expertise, while advanced users can extend the framework to accommodate  
67 custom analysis pipelines. The package has already been applied in published manuscripts and  
68 preprints (Dutta et al., 2025; Pang et al., 2025; Uslan et al., 2025) as well as manuscripts  
69 in preparation, and it has been incorporated into graduate-level training to illustrate how  
70 proteomics workflows parallel to those in single-cell transcriptomics.

71 The applications of `scpviz` span diverse areas of life sciences research, from studying protein  
72 dynamics and signaling pathways in disease models to integrating proteomics with transcrip-  
73 toomics for multi-omics analysis. By bridging the gap between raw mass spectrometry data and  
74 systems-level interpretation, `scpviz` provides a versatile and reproducible platform for advancing  
75 single-cell and spatial proteomics.

## 76 Acknowledgements

77 We acknowledge contributions from Pierre Walker, and support from the A\*STAR Scholarship  
78 (BS-PhD) during the genesis of this project.

79 If you want to cite a software repository URL (e.g. something on GitHub without a preferred  
80 citation) then you can do it with the example BibTeX entry below for Smith et al. (2020).

81 Figures can be included like this: Caption for example figure. and referenced from text using  
82 `section`.

83 Figure sizes can be customized by adding an optional second parameter: Caption for example  
84 figure.

## 85 References

- 86 Dutta, S., Pang, M., Coughlin, G. M., Gudavalli, S., Roukes, M. L., Chou, T.-F., & Grdinaru,  
87 V. (2025, February 11). *Molecularly-guided spatial proteomics captures single-cell identity*

- 88        and heterogeneity of the nervous system. <https://doi.org/10.1101/2025.02.10.637505>
- 89        Pang, M., Jones, J. J., Wang, T.-Y., Quan, B., Kubat, N. J., Qiu, Y., Roukes, M. L., & Chou,  
90        T.-F. (2025). Increasing Proteome Coverage Through a Reduction in Analyte Complexity  
91        in Single-Cell Equivalent Samples. *Journal of Proteome Research*, 24(4), 1528–1538.  
92        <https://doi.org/10.1021/acs.jproteome.4c00062>
- 93        Smith, A. M., Thaney, K., & Hahnel, M. (2020). Fidgit: An ungodly union of GitHub and  
94        figshare. In *GitHub repository*. GitHub. <https://github.com/arfon/fidgit>
- 95        Uslan, T., Quan, B., Wang, T.-Y., Pang, M., Qiu, Y., & Chou, T.-F. (2025). In-Depth  
96        Comparison of Reagent-Based Digestion Methods and Two Commercially Available Kits  
97        for Bottom-Up Proteomics. *ACS Omega*, 10(10), 10642–10652. <https://doi.org/10.1021/acsomega.4c11585>
- 99        Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression  
100      data analysis. *Genome Biology*, 19(1), 15. <https://doi.org/10.1186/s13059-017-1382-0>

DRAFT