The Journal of Open Source Software

scpviz: A Python bioinformatics toolkit for Single-cell Prooteomics-omics analysis

Marion Pang [1¶], **Baiyi Quan** [2], **Ting-Yu Wang** [2], **and Tsui-Fen Chou** [1,2¶]

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

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

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

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 missing values, and are generated within complex experimental designs involving multiple classes of samples (e.g., cell type, treatment, condition). These properties distinguish single-cell proteomics from bulk experiments and create unique challenges in data processing, normalization, and interpretation. The single-cell transcriptomics community has established a mature ecosystem for managing similar challenges, exemplified by the `scanpy` package (Wolf et al., 2018) and the broader `scverse` ecosystem. Building on these foundations, `scpviz` extends the AnnData data structure to the domain of proteomics. It is a Python package that streamlines single-cell and spatial proteomics workflows, supporting a complete analysis pipeline from raw peptide-level data to protein-level summaries and downstream interpretation through differential expression, enrichment analysis, and network analysis. At its core, `scpviz` centers the `pAnnData` class, an `AnnData`-affiliated data structure specialized for proteomics. Together, these components make `scpviz` a comprehensive and extensible framework for single-cell proteomics. By combining flexible data structures, reproducible workflows, and seamless integration with the AnnData/Scanpy ecosystem, the package enables researchers to efficiently connect peptide-level evidence to protein-level interpretation, thereby accelerating methodological development and biological discovery in proteomics.

Statement of need

Although general-purpose data analysis frameworks such as `scanpy` (Wolf et al., 2018) and the broader `scverse` ecosystem have become indispensable for single-cell transcriptomics, comparable tools for proteomics remain limited. Existing proteomics software often focuses on specialized tasks (e.g., peptide identification or spectrum assignment) and does not provide a unified framework for downstream analysis of peptide- and protein-level data within single-cell

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

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

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

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

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

Acknowledgements

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

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

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

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

References

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

and heterogeneity of the nervous system. https://doi.org/10.1101/2025.02.10.637505

Pang, M., Jones, J. J., Wang, T.-Y., Quan, B., Kubat, N. J., Qiu, Y., Roukes, M. L., & Chou, T.-F. (2025). Increasing Proteome Coverage Through a Reduction in Analyte Complexity in Single-Cell Equivalent Samples. *Journal of Proteome Research*, *24*(4), 1528–1538. https://doi.org/10.1021/acs.jproteome.4c00062

Smith, A. M., Thaney, K., & Hahnel, M. (2020). Fidgit: An ungodly union of GitHub and figshare. In *GitHub repository*. GitHub. https://github.com/arfon/fidgit

Uslan, T., Quan, B., Wang, T.-Y., Pang, M., Qiu, Y., & Chou, T.-F. (2025). In-Depth Comparison of Reagent-Based Digestion Methods and Two Commercially Available Kits for Bottom-Up Proteomics. *ACS Omega*, *10*(10), 10642–10652. https://doi.org/10.1021/acsomega.4c11585

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0