The Journal of Open Source Software

scpviz: A Python bioinformatics toolkit for Single-cell Proteomics and multi-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Published: unpublished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 abundance 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 (Virshup et al., 2023). Building on these foundations, scpviz extends the AnnData data structure to the domain of proteomics. `scpviz` 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. The core of `scpviz` is 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` and extended `scverse` 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 focus on specialized tasks (e.g., peptide identification or spectrum assignment) and do not provide a unified framework for downstream analysis of peptide- and protein-level data within single-cell and spatial contexts.

scpviz addresses these gaps by offering an integrated system for the complete proteomics workflow, from raw peptide-level evidence to protein-level summaries and biological interpretation. It is designed for computational biologists and proteomics researchers working with low-input or single-cell datasets from data sources such as Proteome Discoverer or DIA-NN(Demichev et al., 2020).

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 alongside supporting attributes such as .summary, .metadata, .rs matrices (protein–peptide relationships), and .stats. This design allows users to move flexibly between between peptides and proteins while maintaining compatibility with established Python libraries for data science and visualization.

Beyond data organization, scpviz implements proteomics-specific operations, including filtering (e.g., requiring proteins supported by at least two unique peptides), normalization and imputation methods tailored for sparse datasets, and visualization tools such as PCA (Principal Component Analysis), UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction), clustermaps, and abundance plots. For downstream interpretation, it integrates with UniProt for annotation and string-db for enrichment and network analysis (McInnes et al., 2018; Snel et al., 2000; Szklarczyk et al., 2023). The framework also incorporates single-cell proteomics–specific normalization strategies such as directlfq (Ammar et al., 2023), ensuring robust quantification across heterogeneous samples. Finally, pAnnData objects interface seamlessly with scanpy (Wolf et al., 2018) and other ecosystem tools such as harmony (Korsunsky et al., 2019), enabling direct incorporation into established single-cell 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 papers 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 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. The Proteome Exploration Laboratory is partially supported by Caltech Beckman Institute Endowment Funds.

References

Ammar, C., Schessner, J. P., Willems, S., Michaelis, A. C., & Mann, M. (2023). Accurate Label-Free Quantification by directLFQ to Compare Unlimited Numbers of Proteomes. *Molecular & Cellular Proteomics*, *22*(7). https://doi.org/10.1016/j.mcpro.2023.100581

Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, *17*(1), 41–44. https://doi.org/10.1038/s41592-019-0638-x

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

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

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

Snel, B., Lehmann, G., Bork, P., & Huynen, M. A. (2000). STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research*, *28*(18), 3442–3444. https://doi.org/10.1093/nar/28.18.3442

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000

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

Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Berger, B., Pe'er, D., Regev, A., Teichmann, S. A., Finotello, F., Wolf, F. A., Yosef, N., Stegle, O., & Theis, F. J. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology*, *41*(5), 604–606. https://doi.org/10.1038/s41587-023-01733-8

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