

scpviz: A Python bioinformatics toolkit for Single-cell Proteomics and multi-omics analysis

Marion Pang¹, Baiyi Quan², Ting-Yu Wang², 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](https://doi.org/10.xxxxxx/draft)

Software

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

Editor: ¶

Submitted: 16 November 2025

Published: unpublished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

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, 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. Unlike transcriptomic data (counts, reads), proteomics measurements are inherently hierarchical. Peptide-level measurements provide the primary basis from which protein-level quantities are inferred, often with shared or ambiguous

42 peptide assignments and strong reliance on peptide confidence. In addition, missing data
43 is pervasive in low-input and single-cell experiments, further complicating representation and
44 analysis using flat feature-by-sample abstractions.

45 Existing proteomics software typically focuses on upstream tasks such as peptide identification
46 and protein quantification, producing tabular outputs that are not designed for iterative
47 downstream analysis. As a result, users must rely on ad hoc data structures to perform filtering,
48 normalization, visualization, and interpretation, limiting reproducibility and making it difficult
49 to integrate proteomics data into modern single-cell analysis workflows. These limitations
50 are amplified in single-cell and spatial contexts, where complex experimental designs, sparse
51 measurements, and the need to compare across multiple biological conditions require flexible
52 data management and metadata-aware analysis.

53 scpviz addresses this gap by providing a unified framework system for organizing and analyzing
54 proteomics data from raw peptide-level evidence through protein-level summaries and biological
55 interpretation. It is designed for computational biologists and proteomics researchers working
56 with low-input or single-cell datasets generated by common analysis pipelines such as Proteome
57 Discoverer or DIA-NN(Demichev et al., 2020). By enabling structured downstream analysis
58 and integration with established single-cell ecosystems, scpviz supports reproducible and
59 scalable proteomics workflows and facilitates cross-modality analyses that connect protein-level
60 measurements to broader systems-level biology.

61 State of the field

62 A range of software tools exists for proteomics data processing and analysis, each addressing
63 specific stages of the workflow. Upstream platforms such as Proteome Discoverer and DIA-
64 NN provide peptide identification, protein inference, quantitative estimation, and built-in
65 visualization capabilities. However, these environments are primarily designed around fixed
66 analysis pipelines focused on spectral processing and quantification, and offer limited flexibility
67 for downstream data manipulation, including user-defined normalization strategies, imputation
68 methods, and iterative exploratory analysis. As a result, they typically export tabular outputs
69 intended for external analysis rather than serving as extensible frameworks for single-cell or
70 spatial proteomics workflows.

71 In parallel, the single-cell transcriptomics community has developed a mature ecosystem
72 centered on AnnData-based data structures and tools such as scanpy, which support scalable
73 analysis across complex experimental designs. Transcriptomic measurements are naturally
74 represented as gene-level count matrices, enabling downstream analyses to operate on a single
75 level of abstraction. Proteomics analyses, in contrast, center on protein-level interpretation
76 but depend critically on peptide-level measurements for quantification, normalization, and
77 confidence assessment. Existing transcriptomics frameworks do not natively represent this
78 hierarchical evidence structure, limiting their ability to support proteomics-specific operations.

79 With this in mind, scpviz was developed as a standalone framework to bridge this structural
80 gap. Extending upstream proteomics software would not address downstream, metadata-
81 aware analysis needs and is constrained in some cases by commercial software ecosystems,
82 while embedding proteomics-specific logic directly into transcriptomics frameworks would
83 introduce unnecessary complexity for transcriptomics use cases. By extending AnnData with a
84 proteomics-specific data model that explicitly captures peptide-protein relationships, scpviz
85 enables proteomics data to be analyzed within established single-cell workflows while preserving
86 domain-specific rigor. This positions scpviz as connective infrastructure between proteomics
87 and single-cell analysis ecosystems, rather than a replacement for existing tools.

88 Software design

89 The design of `scpviz` centers on the `pAnnData` class, an `AnnData`-affiliated data structure
90 specialized for proteomics. Rather than representing proteomics data as a single flat matrix,
91 `pAnnData` accounts for the hierarchical relationship between peptide-level and protein-level
92 measurements by pairing matched peptide (`.pep`) and protein (`.prot`) `AnnData` objects with
93 supporting attributes such as `.summary`, `.metadata`, `.stats`, and a protein-peptide relationship
94 (`.rs` matrix). This design allows users to preserve explicit peptide-protein relationships during
95 downstream analyses, enabling operations like peptide-level protein abundance normalization
96 or peptide-based fold-change aggregation for differential expression, while maintaining
97 compatibility with established Python libraries for data science and visualization.

98 Proteomics-specific operations in `scpviz` are designed to operate uniformly across peptide
99 and protein-level data. By organizing functionality into mixin-based classes that act on
100 underlying `AnnData` objects, common operations such as filtering, normalization, imputation,
101 and summarization can be applied consistently to both peptides and proteins. This object-
102 oriented design reduces code duplication while ensuring that shared analytical logic respects
103 proteomics-specific constraints at each level of representation. Visualization and analysis
104 utilities (e.g. PCA, UMAP, clustermaps, abundance plots) build directly on these structured
105 representations (McInnes et al., 2018). For downstream interpretation, `scpviz` integrates
106 external resources such as UniProt for annotation and STRING database for functional
107 enrichment and network analysis (Snel et al., 2000; Szklarczyk et al., 2023), and incorporates
108 proteomics-specific quantification strategies such as directLFQ (Ammar et al., 2023). By
109 retaining `AnnData` compatibility, `pAnnData` objects can be used directly with tools such as
110 `scanpy` (Wolf et al., 2018) and `harmony` (Korsunsky et al., 2019), enabling direct incorporation
111 into established single-cell workflows.

112 Design decisions in `scpviz` emphasize separation between data representation and analytical
113 operations. Core data structures encode peptide-protein relationships and metadata, while
114 higher-level methods implement proteomics-aware transformations that can be composed,
115 extended, or replaced. This modular architecture enables users to adapt the framework to
116 evolving experimental designs without coupling downstream analysis logic to upstream data
117 processing assumptions. The design philosophy of `scpviz` thus emphasizes both usability and
118 extensibility. General users can rely on its streamlined API to import, process, and visualize
119 single-cell proteomics data without deep programming expertise, while advanced users can
120 extend the framework to accommodate custom analysis pipelines.

121 Research impact statement

122 `scpviz` has already been used in the analysis of multiple published studies and preprints across
123 single-cell and bulk proteomics applications (Dutta, Pang, Coughlin, et al., 2025; Dutta, Pang,
124 Donahue, et al., 2025; Pang et al., 2025; Uslan et al., 2025). In these works, the framework
125 enabled structured downstream analysis of peptide and protein-level data, including differential
126 expression, functional enrichment, and integration with transcriptomic measurements. In
127 addition, `scpviz` has been incorporated into graduate-level training to demonstrate how
128 proteomics workflows can be analyzed using pipelines common in single-cell transcriptomics
129 analysis, lowering the barrier for new users entering the field.

130 The primary impact of `scpviz` lies in providing reusable infrastructure rather than task-specific
131 analyses. By formalizing peptide-protein relationships within an `AnnData`-compatible data
132 model, the package enables proteomics data to be analyzed using established single-cell tools
133 for visualization, integration, and exploratory analysis. This structured representation supports
134 reproducible downstream workflows and facilitates multi-omics studies in which proteomics
135 and transcriptomics data can be jointly interpreted, enabling analyses of protein abundance
136 dynamics, signaling pathways, and cellular heterogeneity in single-cell and spatial proteomics

137 experiments.

138 scpviz is released as open-source software with comprehensive documentation, automated
139 tests, and reproducible examples, supporting transparent and extensible research workflows.
140 Its emphasis on interoperability with widely used single-cell analysis libraries positions it
141 as infrastructure for continued method development as single-cell and spatial proteomics
142 technologies mature and scale.

143 AI usage disclosure

144 Generative AI tools were used during the development of this work to assist with code
145 refactoring, documentation drafting, and manuscript text editing. All software design decisions,
146 implementation, validation, and scientific interpretation were performed and reviewed by the
147 authors. No generative AI tools were used to generate or analyze research data, and all results
148 reported are reproducible from the publicly available source code and documentation.

149 Acknowledgements

150 We thank Pierre Walker for his many insightful discussions and guidance. We also acknowledge
151 support from the A*STAR BS-PhD Scholarship. The Proteome Exploration Laboratory is
152 partially supported by the Caltech Beckman Institute Endowment Funds.

153 References

- 154 Ammar, C., Schessner, J. P., Willems, S., Michaelis, A. C., & Mann, M. (2023). Accurate
155 Label-Free Quantification by directLFQ to Compare Unlimited Numbers of Proteomes.
156 *Molecular & Cellular Proteomics*, 22(7). <https://doi.org/10.1016/j.mcpro.2023.100581>
- 157 Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-
158 NN: Neural networks and interference correction enable deep proteome coverage in high
159 throughput. *Nature Methods*, 17(1), 41–44. <https://doi.org/10.1038/s41592-019-0638-x>
- 160 Dutta, S., Pang, M., Coughlin, G. M., Gudavalli, S., Roukes, M. L., Chou, T.-F., & Gradinaru,
161 V. (2025, February 11). *Molecularly-guided spatial proteomics captures single-cell identity*
162 *and heterogeneity of the nervous system*. <https://doi.org/10.1101/2025.02.10.637505>
- 163 Dutta, S., Pang, M., Donahue, R. R., Chou, T.-F., Seifert, A. W., & Gradinaru, V. (2025).
164 *Parkinson's disease modeling in regenerative spiny mice (Acomys dimidiatus) captures key*
165 *disease-relevant behavioral, histological, and molecular signatures* (p. 2025.11.06.687049).
166 bioRxiv. <https://doi.org/10.1101/2025.11.06.687049>
- 167 Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner,
168 M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of
169 single-cell data with Harmony. *Nature Methods*, 16(12), 1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>
- 171 McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold
172 Approximation and Projection. *Journal of Open Source Software*, 3(29), 861. <https://doi.org/10.21105/joss.00861>
- 174 Pang, M., Jones, J. J., Wang, T.-Y., Quan, B., Kubat, N. J., Qiu, Y., Roukes, M. L., & Chou,
175 T.-F. (2025). Increasing Proteome Coverage Through a Reduction in Analyte Complexity
176 in Single-Cell Equivalent Samples. *Journal of Proteome Research*, 24(4), 1528–1538.
177 <https://doi.org/10.1021/acs.jproteome.4c00062>
- 178 Snel, B., Lehmann, G., Bork, P., & Huynen, M. A. (2000). STRING: A web-server to retrieve
179 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., Mehryar, 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>

DRAFT