ppx: Programmatic access to proteomics data repositories

William E Fondrie [1], Wout Bittremieux [2,3], and William S Noble [1,4,*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA
[2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[3]Department of Computer Science, University of Antwerp, Antwerp, Belgium
[4]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
*Corresponding Author: william-noble@uw.edu

Abstract

The volume of proteomics and mass spectrometry data available in public repositories continues to grow at a rapid pace as more researchers embrace open science practices. Open access to the data behind scientific discoveries has become critical to validate published findings and develop new computational tools. Here, we present ppx, a Python package that provides easy, programmatic access to the data stored in ProteomeXchange repositories, such as PRIDE and MassIVE. The ppx package can either be used as a command line tool or a Python package to retrieve the files and metadata associated with a project when provided its identifier. To demonstrate how ppx enhances reproducible research, we used ppx within a Snakemake workflow to reanalyze a published dataset with the open modification search tool ANN-SoLo and compared our reanalysis to the original results. We show that ppx readily integrates into workflows and our reanalysis produced results consistent with the original analysis. We envision that ppx will be a valuable tool for creating reproducible analyses, providing tool developers easy access to data for development, testing, and benchmarking, and enabling the use of mass spectrometry data in data-intensive analyses. The ppx package is freely available and open source under the MIT license at: https://github.com/wfondrie/ppx

Introduction

Open access to the data underlying published research is fundamental to foster transparency of the scientific process and to maximize the value of research for the public good. When the data is shared in a findable, accessible, interoperable, and reusable (FAIR) manner, it promotes reproducible research practices and enhances the reliability of the published findings [1, 2]. Data shared under FAIR principles may also yield biological insights beyond the original publication upon secondary analysis with alternative methods. Furthermore, open data provides an invaluable resource for tool developers to refine, test, and benchmark their tools, which ultimately benefits the field.

Fortunately, the proteomics and broader mass spectrometry communities have embraced data sharing and now consistently deposit the raw data behind their publications into public repositories [3]. The ProteomeXchange consortium [4] was formed in 2011 to coordinate the submission and dissemination of mass spectrometry proteomics data worldwide and has seen tremendous growth in the amount of data deposited in its partner repositories—PRIDE [5], PeptideAtlas [6], PASSEL [7], MassIVE [8], jPOST [9], iProx [10], and Panorama Public [11]—since its inception [12]. This push toward open data practices has proven immensely beneficial to the proteomics field, providing consistent datasets for tool development and benchmarking [13–17]. Additionally, the growing abundance of accessible data has been critical for the tools and insights that can only be gained from "big data," including resources such as MassIVE-KB [8] and PRIDE Cluster [18].

Easy, consistent access to the data in these repositories is critical to ensuring the reproducibility of proteomics analyses and leveraging the abundance of data to develop innovative computational approaches.

However, much of this work is currently performed manually; often the task of downloading data is completed by navigating through the repository's web interface to find the file transfer protocol (FTP) link for the specific project of interest. An exception to this pattern is the programmatic access to ProteomeXchange provided by the rpx R package [19].

Here, we present ppx, which provides a simple interface to proteomics repositories both as a command line tool and a Python package. Inspired by rpx, we developed ppx with the goal of improving the reproducibility of proteomics research and offering developers programmatic access to the growing tide of open mass spectrometry data.

Methods and Results

Installation and code availability

The ppx package is available for Python 3.6+ and can be easily installed from the Python Package Index (PyPI) with pip or via conda using the Bioconda channel [20]. The ppx package depends on the requests and tqdm [21] Python packages. As an open source project, ppx is publicly available on GitHub under the permissive MIT license: https://github.com/wfondrie/ppx.

ppx design and implementation

The ppx package is a lightweight Python package that provides a consistent application programming interface (API) to ProteomeXchange and its partner repositories—currently PRIDE and MassIVE—with the goal to enable easy access to the files and metadata associated with each project. When users provide ppx with a ProteomeXchange, PRIDE, or MassIVE identifier, ppx is able to find the partner repository in which the project resides, retrieve metadata about the project, list the files associated with the project, and download the requested files to the user's machine. To accomplish these tasks ppx leverages the metadata provided by the ProteomeXchange XML announcement for a project, as well as the metadata provided by PRIDE or MassIVE. In the case of PRIDE, we specifically use the RESTful API [22] to access information about a project.

We designed ppx to be most effective when project data is stored in a central directory on the user's machine, which is the customizable default for ppx. When a resource is requested, ppx first checks whether the file or metadata has already been downloaded. By default, the remote repositories are only accessed when a new resource is requested. Furthermore ppx adopts a "lazy" approach to fetching data; remote resources are not accessed until they are needed. This behavior ensures that the repository servers are not unnecessarily burdened by requests, allows ppx to remain useful offline, and makes subsequent use of the same resources much faster for the user.

Downloading data via the ppx command line interface is easy. For example, a user can download the FASTA file from PXD000001 [23] using the following command:

```
$ ppx PXD000001 '*.fasta'
```

The ppx Python API offers greater flexibility and can be used to accomplish the same task:

```
>>> import ppx
>>> proj = ppx.find_project('PXD000001')
>>> fasta = proj.remote_files('*.fasta')
>>> local_file = proj.download(fasta)
```

Thus, the files associated with a project are readily accessible from either command line interface or within a user's Python session. Additionally, the ppx Python API provides methods to retrieve the paths to downloaded files and fetch project metadata, making it a powerful tool for data-intensive applications.

ppx is a tool for reproducible research

When attempting to reproduce the results of a published proteomics study, often the first step is to download the raw mass spectrometry data and analyze it in the same manner as the original authors. This has historically been fraught with challenges—insufficient metadata, software installation issues, unpublished in-house code—which renders such exercises impossible. However, a recent trend in the proteomics community has been to embrace tools built for reproducible research. For example, the growth of resources such as Bioconda and BioContainers [25] have made it possible to manage software within virtual environments and containers, relieving the difficulties of installation and managing dependencies. The adoption of workflow engines such as Snakemake

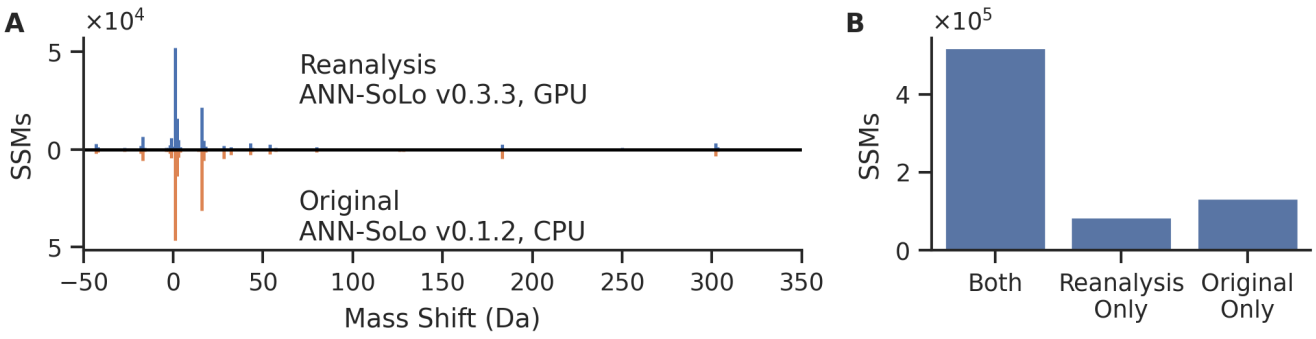


Figure 1: Reanalysis of the Chick et al. [13] HEK293 data with ANN-SoLo. (A) Our reanalysis using ANN-SoLo version 0.3.3 found similar mass shifts for SSMs accepted at 1% FDR when compared to the original analysis [24] conducted with ANN-SoLo version 0.1.2. (B) Although we observed some loss of power, a vast majority of the SSMs from the original analysis were recovered in our reanalysis.

[26], Nextflow [27], and Cromwell [28] provide frameworks to run analysis pipelines reproducibly and orchestrate the necessary environments or containers. Furthermore, workflows built with these engines are readily transferred to the cloud, enabling them to scale massively [29]. Notably, ppx fits well within this framework by providing programmatic access to the data required to execute workflows.

As an example, we used ppx to help reanalyze a fractionated HEK293 cell line dataset [13] with the open modification spectral library search engine ANN-SoLo [24, 30]. Our goal was to reproduce the analysis originally presented by Bittremieux et al. [24], but with an updated version of ANN-SoLo. We used ppx to download the 24 mass spectrometry data files in Mascot generic format (MGF) and the spectral library directly from PRIDE project PXD009861. This library was the MassIVE-KB peptide spectral library (version 2017/11/27), concatenated with decoy spectra that were generated using the shuffle-and-reposition method [31], yielding 3,009,902 total spectra. We searched this data using ANN-SoLo version 0.3.3, as opposed to ANN-SoLo version 0.1.2 which was used for the original analysis. Notably, the newer version of ANN-SoLo supports the use of graphics processing units (GPUs) to massively speed up the search, at the cost of a slight loss statistical power to detect peptides [30]. Correspondingly, we chose search parameters that best matched the original ANN-SoLo analysis of the HEK293 data for our searches. Additionally, we set the hash length to 400 and used GPUs to accelerate our searches, leaving other parameters to their defaults.

After running our analysis, we compared the spectrum-spectrum matches (SSMs) detected at a 1% false discovery rate (FDR) in our reanalysis against the original results that were uploaded to the PRIDE project alongside the raw data. We found that the differences in precursor mass—or "mass shifts"—found in our reanalysis closely matched those that were originally reported (Figure 1A). As expected, we did observe a small loss of power in the total number of detected SSMs when using the GPU (Figure 1B); however, we also found that both the unmodified SSMs and those bearing the most common mass shifts were largely consistent between the analyses.

Critically, every step of our reanalysis—downloading the raw data from PRIDE with ppx, searching with ANN-SoLo, downloading the original analysis results from PRIDE with ppx, and the creation of our figures with Python—is fully encapsulated in a Snakemake workflow that is publicly available at `https://github.com/Noble-Lab/ppx-workflow`. We chose to parallelize this analysis on our high-performance computing cluster, equipped with 12 NVIDIA RTX 2080 GPUs. However, the same workflow can be readily configured to reproduce our results on a single machine or in the cloud, provided sufficient computational resources are available.

Conclusions

Here, we have introduced ppx and provided a short vignette on how ppx can be used in combination with a workflow engine to create fully reproducible analyses. The ppx package is a powerful tool for enabling reproducible research, and we envision that it will be particularly useful for tool developers to access the mountains of public mass spectrometry data at their dis-

posal. Furthermore, the easy programmatic access to data that ppx provides will present great opportunities for the development of innovative approaches that require big data to be effective. Finally, we anticipate that ppx will become more valuable as more mass spectrometry experiments are annotated with standardized metadata [32], which lays the groundwork for fully automated analysis pipelines.

Acknowledgments

The research reported in this publication was supported by the National Institutes of Health awards T32HG000035, P41GM103533, and R01GM121818. W.B. is a postdoctoral researcher of the Research Foundation – Flanders (FWO 12W0418N).

References

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” eng. In: *Scientific Data* 3 (Mar. 2016), p. 160018.

[2] Wilson, S. L., Way, G. P., Bittremieux, W., Armache, J.-P., et al. “Sharing Biological Data: Why, When, and How.” eng. In: *FEBS letters* 595.7 (Apr. 2021), pp. 847–863.

[3] Martens, L. and Vizcaíno, J. A. “A Golden Age for Working with Public Proteomics Data.” eng. In: *Trends in Biochemical Sciences* 42.5 (May 2017), pp. 333–341.

[4] Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., et al. “ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination.” eng. In: *Nature Biotechnology* 32.3 (Mar. 2014), pp. 223–226.

[5] Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., et al. “The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data.” eng. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D442–D450.

[6] Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., et al. “The PeptideAtlas Project.” eng. In: *Nucleic Acids Research* 34.Database issue (Jan. 2006), pp. D655–658.

[7] Farrah, T., Deutsch, E. W., Kreisberg, R., Sun, Z., et al. “PASSEL: The PeptideAtlas SRM Experiment Library.” In: *PROTEOMICS* 12.8 (Feb. 9, 2012), pp. 1170–1175.

[8] Wang, M., Wang, J., Carver, J., Pullman, B. S., et al. “Assembling the Community-Scale Discoverable Human Proteome.” eng. In: *Cell Systems* 7.4 (Oct. 2018), 412–421.e5.

[9] Watanabe, Y., Yoshizawa, A. C., Ishihama, Y., and Okuda, S. “The jPOST Repository as a Public Data Repository for Shotgun Proteomics.” eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 2259 (2021), pp. 309–322.

[10] Ma, J., Chen, T., Wu, S., Yang, C., et al. “iProX: An Integrated Proteome Resource.” In: *Nucleic Acids Research* 47.D1 (Jan. 8, 2019), pp. D1211–D1217.

[11] Sharma, V., Eckels, J., Schilling, B., Ludwig, C., et al. “Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline.” eng. In: *Molecular & cellular proteomics: MCP* 17.6 (June 2018), pp. 1239–1244.

[12] Deutsch, E. W., Bandeira, N., Sharma, V., Perez-Riverol, Y., et al. “The ProteomeXchange Consortium in 2020: Enabling ’big Data’ Approaches in Proteomics.” In: *Nucleic Acids Research* (Nov. 5, 2019).

[13] Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., et al. “A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides.” eng. In: *Nature Biotechnology* 33.7 (July 2015), pp. 743–749.

[14] Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., et al. “A Multicenter Study Benchmarks Software Tools for Label-Free Proteome Quantification.” eng. In: *Nature Biotechnology* 34.11 (Nov. 2016), pp. 1130–1136.

[15] Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., et al. “Building ProteomeTools Based on a Complete Synthetic Human Proteome.” eng. In: *Nature Methods* 14.3 (Mar. 2017), pp. 259–262.

[16] Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., et al. “Mass-Spectrometry-Based Draft of the Human Proteome.” eng. In: *Nature* 509.7502 (May 2014), pp. 582–587.

[17] Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., et al. “A Draft Map of the Human Proteome.” eng. In: *Nature* 509.7502 (May 2014), pp. 575–581.

[18] Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D. L., et al. “Recognizing Millions of Consistently Unidentified Spectra across Hundreds of Shotgun Proteomics Datasets.” eng. In: *Nature Methods* 13.8 (Aug. 2016), pp. 651–656.

[19] Gatto, L. *Rpx*. Bioconductor. 2017.

[20] Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., et al. “Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences.” en. In: *Nature Methods* 15.7 (July 2018), pp. 475–476.

[21] Casper da Costa-Luis, Larroque, S. K., Altendorf, K., Mary, H., et al. *Tqdm: A Fast, Extensible Progress Bar for Python and CLI*. Zenodo. Apr. 2021.

[22] Reisinger, F., del-Toro, N., Ternent, T., Hermjakob, H., et al. “Introducing the PRIDE Archive RESTful Web Services.” eng. In: *Nucleic Acids Research* 43.W1 (July 2015), W599–604.

[23] Gatto, L. and Christoforou, A. “Using R and Bioconductor for Proteomics Data Analysis.” eng. In: *Biochimica Et Biophysica Acta* 1844.1 Pt A (Jan. 2014), pp. 42–51.

[24] Bittremieux, W., Meysman, P., Noble, W. S., and Laukens, K. “Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing.” eng. In: *Journal of Proteome Research* 17.10 (Oct. 2018), pp. 3463–3474.

[25] Leprevost, F. V., Grüning, B. A., Alves Aflitos, S., Röst, H. L., et al. “BioContainers: An Open-Source and Community-Driven Framework for Software Standardization.” eng. In: *Bioinformatics (Oxford, England)* 33.16 (Aug. 2017), pp. 2580–2582.

[26] Köster, J. and Rahmann, S. “Snakemake–a Scalable Bioinformatics Workflow Engine.” eng. In: *Bioinformatics (Oxford, England)* 28.19 (Oct. 2012), pp. 2520–2522.

[27] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., et al. “Nextflow Enables Reproducible Computational Workflows.” eng. In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319.

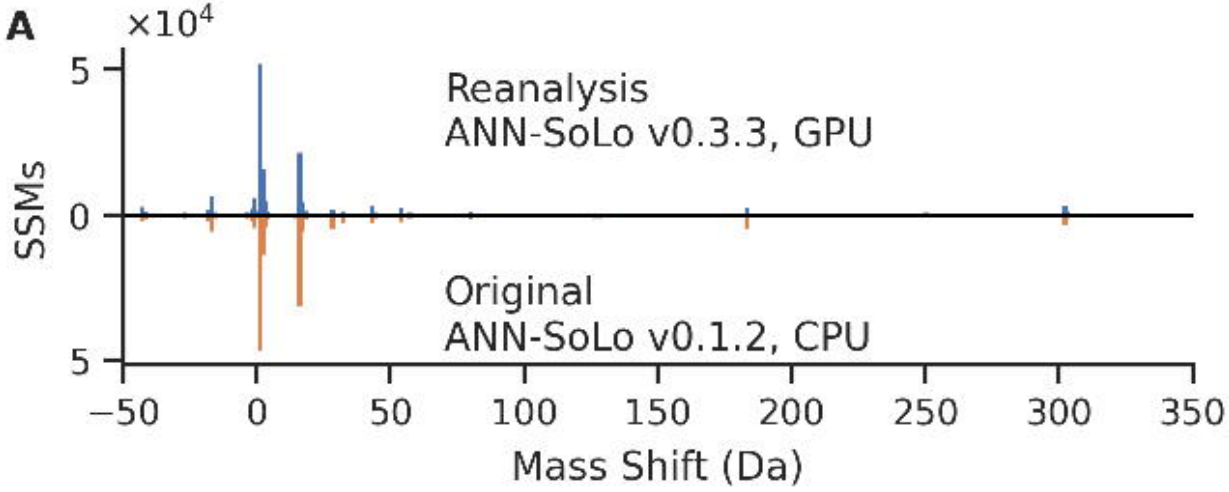
[28] Voss, K., Auwera, G. V. D., and Gentry, J. *Full-Stack Genomics Pipelining with GATK4 + WDL + Cromwell [Version 1; Not Peer Reviewed]*. Vol. 6. F1000Research, 2017.

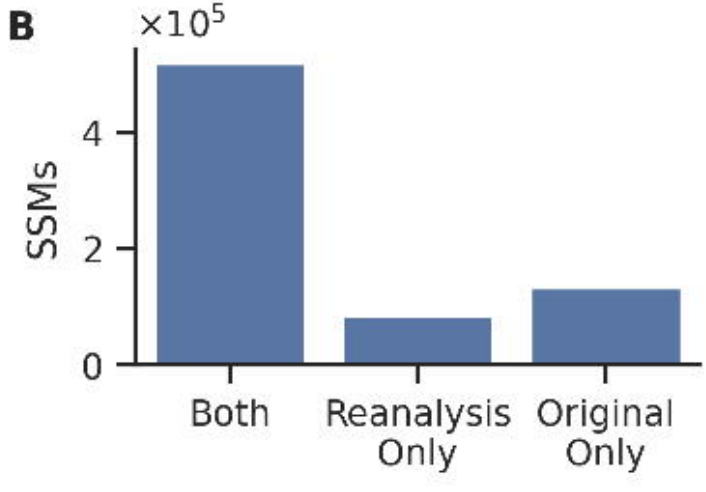
[29] Neely, B. A. “Cloudy with a Chance of Peptides: Accessibility, Scalability, and Reproducibility with Cloud-Hosted Environments.” eng. In: *Journal of Proteome Research* 20.4 (Apr. 2021), pp. 2076–2082.

[30] Bittremieux, W., Laukens, K., and Noble, W. S. “Extremely Fast and Accurate Open Modification Spectral Library Searching of High-Resolution Mass Spectra Using Feature Hashing and Graphics Processing Units.” eng. In: *Journal of Proteome Research* 18.10 (Oct. 2019), pp. 3792–3799.

[31] Lam, H., Deutsch, E. W., and Aebersold, R. “Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics.” eng. In: *Journal of Proteome Research* 9.1 (Jan. 2010), pp. 605–610.

[32] Dai, C., Fullgrabe, A., Pfeuffer, J., Solovyeva, E., et al. “A Proteomics Sample Metadata Representation for Multiomics Integration, and Big Data Analysis.” en. In: *bioRxiv* (May 2021), p. 2021.05.21.445143.


A
×10^4
5
0
5
SSMs
Reanalysis
ANN-SoLo v0.3.3, GPU
Original
ANN-SoLo v0.1.2, CPU
−50
0
50
100
150
200
250
300
350
Mass Shift (Da)


B
×10^5
4
2
0
SSMs
Both
Reanalysis Only
Original Only