Received: 18 October 2023 | Accepted: 24 October 2023

DOI: 10.1002/pmic.202300354

Proteomics
Proteomics and Systems Biology

COMMENTARY

The need for open and FAIR data in top-down proteomics

Frederik Lermyte

Department of Chemistry, Clemens-Schöpf Institute of Organic Chemistry and Biochemistry, Technical University of Darmstadt, Darmstadt, Germany

Correspondence
Frederik Lermyte, Department of Chemistry, Clemens-Schöpf Institute of Organic Chemistry and Biochemistry, Technical University of Darmstadt,
Peter-Grünberg-Straße 4, 64287 Darmstadt, Germany.
Email: frederik.lermyte@tu-darmstadt.de

Abstract

In recent years, there has been a tremendous evolution in the high-throughput, tandem mass spectrometry-based analysis of intact proteins, also known as top-down proteomics (TDP). Both hardware and software have developed to the point that the technique has largely entered the mainstream, and large-scale, ambitious, multi-laboratory initiatives have started to make their appearance in the literature. For this, however, more convenient and robust data sharing and reuse will be required. Walzer et al. have created TopDownApp, a customisable, open platform for visualisation and analysis of TDP data, which they hope will be a step in this direction. As they point out, other benefits of such data sharing and interoperability would include reanalysis of published datasets, as well as the prospect of using large amounts of data to train machine learning algorithms. In time, this work could prove to be a valuable resource in the move towards a future of greater TDP data findability, accessibility, interoperability and reusability.

KEYWORDS
FAIR data, mass spectrometry, open science, top-down proteomics

In this issue of *PROTEOMICS*, Walzer et al. [1] present TopDownApp, an open-source platform for analysis and visualisation of top-down proteomics (TDP) data. In addition to the tool itself, they make the case that increased data sharing—for which they propose TopDownApp can play a role—is crucial for the future of the field.

TDP was first highlighted as a method to watch in Nature Methods in 2008 [2]. At the time, it was pointed out that the field was still rather focussed on the analysis of individual, purified proteins—something arguably more accurately referred to as top-down protein mass spectrometry (MS) rather than true TDP [3]. Another issue that was pointed out in 2008 was the relative lack of robust software tools [2]. In subsequent years, significant improvements were made to all aspects of the top-down experiment, crucially including separation. This led to the analysis of more complex, biologically relevant samples, including a landmark study in 2011 in which top-down MS was applied on a proteome scale to human cells [4].

As separation, fragmentation and mass analyser performance improved over the years, there was a greater need for reliable software for TDP. As such, bioinformatics tools evolved, and currently there exists a wide array of vendor-specific or -neutral software packages, with some being freely available and others being commercial in nature. Some of the more prominent options were presented in References [5–11], and Schaffer et al. recently reviewed the field [12]. The Consortium for TDP maintains a list of useful software packages at https://www.topdownproteomics.org/resources/software/. The vast improvements in both hardware and software in recent years have led to an ever-increasing number of studies applying top-down MS to study both primary and higher-order protein structure [13, 14].

As individual laboratories have become more experienced and leveraged the aforementioned technological improvements to become more efficient at carrying out top-down studies, there has been an increasing trend towards collaboration. The Consortium for TDP was

founded in 2012, and was a key driving force behind the introduction of the now widely adopted term 'proteoform' [15]. In 2014, an inter-laboratory pilot project was carried out for the top-down characterisation of histone proteoforms [16]. More recently, the same consortium has published guidelines for intact protein mass measurements and top-down analysis [17], and also carried out an inter-laboratory study for the top- and middle-down characterisation of monoclonal antibodies [18].

The field is evolving towards increased collaboration, which will be essential to tackle some of the ambitious projects that have recently been proposed, for example the development of tissue- and pathology-specific proteoform atlases [19–21]. The recently proposed Human Proteoform Project should be highlighted in particular, as it has the ambition to ultimately rival the scope of the Human Genome Project [22]. As such, seamless collaboration and communication between laboratories across institutes, nations and continents will be essential. As much of the data in the TDP field currently exists in workflow- and vendor-specific 'silos', a more profound development of vendor-neutral, ideally open-source software will therefore be needed, as will a shift (both technological and cultural) towards sharing data according to FAIR principles [23]. The work by Walzer et al. [1] in this issue attempts to provide a step in this direction. The TopDownApp presented in this work makes heavy use of open formats like mzML and mzTab, including for the reporting of outputs.

One interesting aspect of TopDownApp is the emphasis on deconvolution. Knowledge of the intact mass of a protein is a key benefit of TDP; therefore, high-quality deconvolution of precursor spectra is especially critical [24]. Walzer et al. [1] have developed a convenient visualisation in which observed isotope clusters at different charge states are aligned on a common mass (not *m/z*) axis, which allows a user to rapidly assess the quality of the deconvolution result. In another section of the study, the authors use several different proteoform identification software methods to reanalyse a publicly available dataset associated with the original Blood Proteoform Atlas publication [19], and note that the choice of method has a significant effect on the number of identified proteoforms [24].

One limitation of TopDownApp is that currently, only one vendor-specific format can be read, specifically Thermo RAW files, while for other vendors, users have to convert the data to mzML first. While conversion tools are available, this introduces a (small) obstacle for potential users. It can be hoped that this limitation will be remedied in the future, and the authors explicitly anticipate plugging in converters for other vendor-specific formats if and when they become available [1]. In general, some experimentalists might perceive a barrier to entry in the fairly substantial degree of bioinformatics proficiency that seems to be required to make full use of TopDownApp. Alternatively, this could be interpreted as merely a sign that closer collaboration with bioinformaticians might be an intrinsic part of the future of the TDP field. Possibly, more user-friendly iterations of this or other software packages might arise and could lead to improved adoption, although one could imagine that there might be trade-offs in flexibility and capability of such versions.

The work by Walzer et al. [1] highlights the need for a cultural shift in the TDP field, as they point out that only a few hundred TDP datasets are available in the PRIDE repository (out of a total of more than 35,000 datasets) [25, 26]. This discrepancy cannot be fully attributed to the greater number of bottom-up studies that have been performed compared to top-down, and certainly many more than a few hundred datasets have been acquired over the years associated with published studies. Clearly, as practitioners of TDP, there is scope for us to do better. In this context, only time will tell whether or not TopDownApp will turn out to be a critical tool that the TDP community has been waiting for. Zooming out though, this work should be seen foremost as a potentially important step and call to action to make TDP more open and FAIR, in order to tackle the monumentally ambitious projects that are envisioned to be undertaken by this community in the coming years.

ACKNOWLEGEMENTS

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT

Not applicable.

ORCID

Frederik Lermyte https://orcid.org/0000-0001-7371-4475

REFERENCES

1. Walzer, M., Jeong, K., Tabb, D. L., & Vizcaino, J. A. (2023). TopDownApp: An open and modular platform for analysis and visualisation of top-down proteomics data. *Proteomics*, e2200403.
2. Doerr, A. (2008). Top-down mass spectrometry. *Nature Methods*, *5*, 24.
3. Lermyte, F., Tsybin, Y. O., O'connor, P. B., & Loo, J. A. (2019). Top or middle? Up or down? Toward a standard lexicon for protein top-down and allied mass spectrometry approaches up or down? Toward a standard lexicon for protein top-down and allied mass spectrometry approaches. *Journal of the American Society for Mass Spectrometry*, *30*, 1149–1157. https://doi.org/10.1007/s13361-019-02201-x
4. Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E, Catherman, A. D., Durbin, K. R., Tipton, J. D., Vellaichamy, A., Kellie, J. F., Li, M., Wu, C., Sweet, S. M. M., Early, B. P., Siuti, N., Leduc, R. D., Compton, P. D., Thomas, P. M., & Kelleher, N. L. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, *480*, 254–258. https://doi.org/10.1038/nature10575
5. Larson, E. J., Pergande, M. R., Moss, M. E., Rossler, K. J., Wenger, R. K., Krichel, B., Josyer, H., Melby, J. A., Roberts, D. S., Pike, K., Shi, Z., Chan, H.-J., Knight, B., Rogers, H. T., Brown, K. A., Ong, I. M., Jeong, K., Marty, M. T., Mcilwain, S. J., & Ge, Y. (2023). MASH native: A unified solution for native top-down proteomics data processing. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btad359
6. Wu, Z., Roberts, D. S., Melby, J. A., Wenger, K., Wetzel, M., Gu, Y., Ramanathan, S. G., Bayne, E. F., Liu, X., Sun, R., Ong, I. M., Mcilwain, S. J., & Ge, Y. (2020). MASH Explorer: A universal software environment for top-down proteomics. *Journal of Proteome Research*, *19*, 3867–3876. https://doi.org/10.1021/acs.jproteome.0c00469
7. Kou, Q., Xun, L., & Liu, X. (2016). TopPIC: A software tool for top-down mass spectrometry-based proteoform identification and

characterization. *Bioinformatics*, *32*, 3495–3497. https://doi.org/10.1093/bioinformatics/btw398
8. Zamdborg, L., Leduc, R. D., Glowacz, K. J., Kim, Y.-B., Viswanathan, V., Spaulding, I. T., Early, B. P., Bluhm, E. J., Babai, S., & Kelleher, N. L. (2007). ProSight PTM 2.0: Improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids*, *35*, W701–W706. https://doi.org/10.1093/nar/gkm371
9. Fellers, R. T., Greer, J. B., Early, B. P., Yu, X., LeDuc, R. D., Kelleher, N. L., & Thomas, P. M. (2015). Cover picture: Proteomics 7'15. *Proteomics*, *15*, 1235–1238. https://doi.org/10.1002/pmic.201570050
10. Solntsev, S. K., Shortreed, M. R., Frey, B. L., & Smith, L. M. (2018). Enhanced global post-translational modification discovery with MetaMorpheus. *Journal of Proteome Research*, *17*, 1844–1851. https://doi.org/10.1021/acs.jproteome.7b00873
11. Lantz, C., Zenaidee, M. A., Wei, B., Hemminger, Z., Ogorzalek Loo, R. R., & Loo, J. A. (2021). ClipsMS: An algorithm for analyzing internal fragments resulting from top-down mass spectrometry. *Journal of Proteome Research*, *20*, 1928–1935. https://doi.org/10.1021/acs.jproteome.0c00952
12. Schaffer, L. V., Millikin, R. J., Miller, R. M., Anderson, L. C., Fellers, R. T., Ge, Y., Kelleher, N. L., Leduc, R. D., Liu, X., Payne, S. H., Sun, L., Thomas, P. M., Tucholski, T., Wang, Z., Wu, S., Wu, Z., Yu, D., Shortreed, M. R., & Smith, L. M. (2019). Identification and quantification of proteoforms by mass spectrometry. *Proteomics*, *19*, e1800361. https://doi.org/10.1002/pmic.201800361
13. Zhou, M., Lantz, C., Brown, K. A., Ge, Y., Pasa-Tolic, L., Loo, J. A., & Lermyte, F. (2020). Higher-order structural characterisation of native proteins and complexes by top-down mass spectrometry. *Chemical Science*, *11*, 12918–12936. https://doi.org/10.1039/d0sc04392c
14. Habeck, T., & Lermyte, F. (2023). Seeing the complete picture: Proteins in top-down mass spectrometry. *Essays in Biochemistry*, *67*, 283–300. https://doi.org/10.1042/EBC20220098
15. Smith, L. M., & Kelleher, N. L. (2013). Proteoform: A single term describing protein complexity. *Nature Methods*, *10*, 186–187. https://doi.org/10.1038/nmeth.2369
16. Dang, X., Scotcher, J., Wu, S., Chu, R. K., Tolic, N., Ntai, I., Thomas, P. M., Fellers, R. T., Early, B. P., Zheng, Y., Durbin, K. R., Leduc, R. D., Wolff, J. J., Thompson, C. J., Pan, J., Han, J., Shaw, J. B., Salisbury, J. P., Easterling, M., & Young, N. L. (2014). The first pilot project of the consortium for top-down proteomics: A status report. *Proteomics*, *14*, 1130–1140. https://doi.org/10.1002/pmic.201300438
17. Donnelly, D. P., Rawlins, C. M., Dehart, C. J., Fornelli, L., Schachner, L. F., Lin, Z., Lippens, J. L., Aluri, K. C., Sarin, R., Chen, B., Lantz, C., Jung, W., Johnson, K. R., Koller, A., Wolff, J. J., Campuzano, I. D. G., Auclair, J. R., Ivanov, A. R., Whitelegge, J. P., … Agar, J. N. (2019). Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nature Methods*, *16*, 587–594. https://doi.org/10.1038/s41592-019-0457-0
18. Srzentic, K., Fornelli, L., Tsybin, Y. O., Loo, J. A., Seckler, H., Agar, J. N., Anderson, L. C., Bai, D. L., Beck, A., Brodbelt, J. S., Van Der Burgt, Y. E. M., Chamot-Rooke, J., Chatterjee, S., Chen, Y., Clarke, D. J., Danis, P. O., Diedrich, J. K., D'ippolito, R A., Dupré, M., … Zhou, M. (2020). Interlaboratory study for characterizing monoclonal antibodies by top-down and middle-down mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *31*, 1783–1802. https://doi.org/10.1021/jasms.0c00036
19. Melani, R. D., Gerbasi, V. R., Anderson, L. C., Sikora, J. W., Toby, T. K., Hutton, J. E., Butcher, D. S., Negrão, F., Seckler, H. S., Srzentic, K., Fornelli, L., Camarillo, J. M., Leduc, R. D., Cesnik, A J., Lundberg, E., Greer, J. B., Fellers, R. T., Robey, M. T., Dehart, C. J., … Kelleher, N. L. (2022). The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. *Science*, *375*, 411–418. https://doi.org/10.1126/science.aaz5284
20. Drown, B. S., Jooß, K., Melani, R. D., Lloyd-Jones, C., Camarillo, J. M., & Kelleher, N. L. (2022). Mapping the proteoform landscape of five human tissues. *Journal of Proteome Research*, *21*, 1299–1310. https://doi.org/10.1021/acs.jproteome.2c00034
21. Hollas, M. A. R., Robey, M. T., Fellers, R. T., Leduc, R. D., Thomas, P. M., & Kelleher, N. L. (2022). The Human Proteoform Atlas: A FAIR community resource for experimentally derived proteoforms. *Nucleic Acids Research*, *50*, D526–D533. https://doi.org/10.1093/nar/gkab1086
22. Smith, L. M., Agar, J. N., Chamot-Rooke, J., Danis, P. O., Ge, Y., Loo, J. A., Pasa-Tolic, L., Tsybin, Y. O., & Kelleher, N. L. (2021). The Human Proteoform Project: Defining the human proteome. *Science Advances*, *7*, eabk0734. https://doi.org/10.1126/sciadv.abk0734
23. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18
24. Tabb, D. L., Jeong, K., Druart, K., Gant, M. S., Brown, K. A., Nicora, C., Zhou, M., Couvillion, S., Nakayasu, E., Williams, J. E., Peterson, H. K., Mcguire, M. K., Mcguire, M. A., Metz, T. O., & Chamot-Rooke, J. (2023). Comparing top-down proteoform identification: Deconvolution, PrSM overlap, and PTM detection. *Journal of Proteome Research*, *22*, 2199–2217. https://doi.org/10.1021/acs.jproteome.2c00673
25. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D. J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., Walzer, M., Wang, S., Brazma, A., & Vizcaíno, J. A. (2022). The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, *50*, D543–D552. https://doi.org/10.1093/nar/gkab1038
26. Deutsch, E. W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J. J., Kundu, D. J., García-Seisdedos, D., Jarnuczak, A. F., Hewapathirana, S., Pullman, B. S., Wertz, J., Sun, Z., Kawano, S., Okuda, S., Watanabe, Y., Hermjakob, H., Maclean, B., Maccoss, M. J., Zhu, Y., … Vizcaíno, J. A. (2020). The ProteomeXchange consortium in 2020: Enabling 'big data' approaches in proteomics. *Nucleic Acids Research*, *48*, D1145–D1152. https://doi.org/10.1093/nar/gkz984

How to cite this article: Lermyte, F. (2024). The need for open and FAIR data in top-down proteomics. *Proteomics*, *24*, e2300354. https://doi.org/10.1002/pmic.202300354