ProteomicsML: an online platform for community-curated data sets and tutorials for machine learning in proteomics

Rehfeldt, T.G.; Gabriels, R.; Bouwmeester, R.; Gessulat, S.; Neely, B.A.; Palmblad, M.; ... ; Deutsch, E.W.

Citation
Rehfeldt, T. G., Gabriels, R., Bouwmeester, R., Gessulat, S., Neely, B. A., Palmblad, M., ... Deutsch, E. W. (2023). ProteomicsML: an online platform for community-curated data sets and tutorials for machine learning in proteomics. *Journal Of Proteome Research*, *22*(2), 632-636. doi:10.1021/acs.jproteome.2c00629

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3593850

Note: To cite this publication please use the final published version (if applicable).

Journal of proteome research

pubs.acs.org/jpr

Technical Note

ProteomicsML: An Online Platform for Community-Curated Data sets and Tutorials for Machine Learning in Proteomics

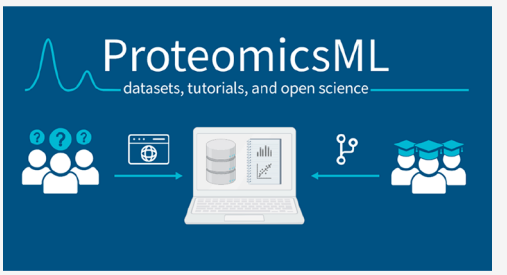
Tobias G. Rehfeldt,[+] Ralf Gabriels,[+] Robbin Bouwmeester,[+] Siegfried Gessulat, Benjamin A. Neely, Magnus Palmblad, Yasset Perez-Riverol, Tobias Schmidt, Juan Antonio Vizcaíno,* and Eric W. Deutsch*

Cite This: *J. Proteome Res.* 2023, 22, 632−636

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Data set acquisition and curation are often the most difficult and time-consuming parts of a machine learning endeavor. This is especially true for proteomics-based liquid chromatography (LC) coupled to mass spectrometry (MS) data sets, due to the high levels of data reduction that occur between raw data and machine learning-ready data. Since predictive proteomics is an emerging field, when predicting peptide behavior in LC-MS setups, each lab often uses unique and complex data processing pipelines in order to maximize performance, at the cost of accessibility and reproducibility. For this reason we introduce ProteomicsML, an online resource for proteomics-based data sets and tutorials across most of the currently explored physicochemical peptide properties. This community-driven resource makes it simple to access data in easy-to-process formats, and contains easy-to-follow tutorials that allow new users to interact with even the most advanced algorithms in the field. ProteomicsML provides data sets that are useful for comparing state-of-the-art machine learning algorithms, as well as providing introductory material for teachers and newcomers to the field alike. The platform is freely available at https://www.proteomicsml.org/, and we welcome the entire proteomics community to contribute to the project at https://github.com/ProteomicsML/ProteomicsML.



KEYWORDS: *machine learning, deep learning, proteomics, educational platform, community platform, bioinformatics*

Downloaded via LEIDEN UNIV on May 30, 2023 at 13:50:35 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

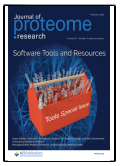
Computational predictions of analyte behavior in the context of mass spectrometry (MS) data have been explored for nearly five decades, with early rudimentary predictions dating back to 1983.[1] With the rise of technology and computational power, machine learning (ML) approaches were introduced into the field of proteomics in 1998[2] and ML-based models quickly overtook human accuracy. Since then, dozens of articles have described efforts to train models for a multitude of physicochemical properties associated with the field of high-throughput proteomics, as reviewed by Neely et al.[3] Some of the most-commonly studied properties are retention time and fragmentation spectrum intensities, while a large range of lesser explored properties exists as well. For an exhaustive review of the current undertakings, see Wen et al. and Bouwmeester et al.[4,5] While many of these efforts are still in the realm of basic exploratory research, ML approaches are increasingly being incorporated into mainstream tools and standalone predictive resources.[4,6−8]

When training any ML model, it is crucial to obtain suitable training and evaluation data sets. Likewise, in many fields of research where ML is applied, it is common to have a range of educational data sets, such as the MNIST (Modified National Institute of Standards and Technology)[9] or IRIS (https://archive.ics.uci.edu/ml/datasets/iris) data sets, allowing newcomers to the field to easily learn common ML methodologies. Likewise, state-of-the-art models can use benchmark data sets such as ImageNet (https://www.image-net.org) or those available on the UCI Machine Learning Repository (https://archive.ics.uci.edu) to compare their predictive capabilities. Similar to the utility of benchmark data sets, such as the number of survivors on the Titanic, which has been modeled more than 54 000 times (https://www.kaggle.com/competitions/titanic), we seek to define proteomics data sets that can provide an entry point for ML modeling.

Although there have been numerous efforts to explore the predictive capabilities of models, there are barriers that limit widespread adoption in the field of predictive proteomics. First, there are considerable difficulties in accessing data sets in a suitable form for ML applications. A substantial effort is required to prepare raw proteomics data sets into a format usable for ML, as this demands extensive knowledge of the multitude of proteomics file formats and postprocessing methods. MS data also has a tendency to be fraught with missing metadata, making it challenging to compare across data sets. Furthermore, most

Special Issue: Software Tools and Resources 2023

Received: October 4, 2022
Published: January 24, 2023



ACS Publications

© 2023 The Authors. Published by American Chemical Society

ML frameworks in proteomics implement dedicated post-processing pipelines to prepare the files for ML algorithms. Recently, tools such as ppx[10] and MS2AI[11] were created to facilitate this process, but they are still limited to certain use cases due to the complex nature of liquid chromatography coupled to mass spectrometry (LC-MS) data.

Second, while some ML-ready data sets are available on platforms such as Kaggle[12] or in supplementary tables of publications, they are often difficult to find and lack long-term maintenance and support postpublication. While there is no formal consensus in the field, there are certain data sets that are often used for training such as ProteomeTools.[13] Nevertheless, there are no widely used data sets used to compare the performance of tools developed by different researchers, making it difficult for new algorithms to be evaluated and compared to older tools. This issue is only further exacerbated by individual groups relying on different pre- and postprocessing protocols, such as differences in normalization of measurements or in the implementation of model performance metrics.

As an outcome of the 2022 Lorentz Center Workshop on Proteomics and Machine Learning (Leiden, The Netherlands, March 2022), we have created a web platform to facilitate the application of ML approaches to the field of MS-based proteomics. The resource is intended to provide a central focal point for curating and disseminating data sets that are ready to use for ML research, and to encourage new entrants into the field through expert-driven tutorials.

Here we describe how ProteomicsML has been developed using commonly available tools and designed for future ease of maintenance. We provide a brief overview of the data sets that are currently available at ProteomicsML and how it can be expanded in the future with more data. We also describe the initial set of tutorials that can be used as an introduction to the field of ML in proteomics.

THE PROTEOMICSML PLATFORM

The primary entry point for the resource is the ProteomicsML Web site (https://www.proteomicsml.org/). It contains general introductory data sets that are already preprocessed and ready for training or evaluation, and contains educational resources in the form of tutorials for those new to ML in proteomics. The code base for the Web site is maintained via a GitHub repository (https://github.com/ProteomicsML/ProteomicsML), and is therefore easy to maintain and amenable to outside contributions from the community. On the GitHub repository, researchers can open pull requests (proposals for adding or changing information) for new data sets or tutorials. These pull requests are then reviewed by the maintainers, currently the authors of this paper, in line with the guidelines in the contributing section of the ProteomicsML Web site. Data sets and tutorials hosted as part of the GitHub repository fall under the CC BY 4.0 license, as indicated on both the repository and the Web site. The PRIDE database infrastructure[14] is also used to store larger data sets on an FTP server dedicated to ProteomicsML.

A key goal of ProteomicsML is to advance with the field, which is why we provide a platform with detailed documentation, including a contributing guide on how to upload data sets and tutorials for specific ML workflows or algorithms. After curation by the maintainers, the contributions have to pass a build test in order to maintain integrity of the platform, and, if passed, are automatically published on the Web site and are freely accessible to other researchers.

For many LC-MS properties, such as retention time and fragmentation intensity, well-performing ML models have already been published. We aim to provide suitable data sets and tutorials to easily reproduce these results in an educational fashion. All data sets on the platform are organized by data type, and should ideally be provided in a simple data format that is suitable for direct import into ML toolkits. Each data type can contain one or more data sets for different purposes, and each data set should be sufficiently annotated with metadata (e.g., its origin, how it was processed, and the relevant literature citations).

Along with well-annotated data sets, the platform provides users with in-depth tutorials on how to download, import, handle, and train various ML models. Many of the LC-MS data types require certain, sometimes complex, preprocessing steps in order to be fully compatible with ML frameworks. For this reason, we believe it is crucial to provide guidelines on these processes to ultimately lower the entry barriers for new users to the field. Tutorials on ProteomicsML can be attribute- or data set-specific, allowing new tutorial submissions to focus on either the direct interactions with specific ML models or methodologies, or on a certain aspect of data preprocessing.

Often when new modeling approaches are published, they are accompanied by data sets with novel pre- and postprocessing steps. Using ProteomicsML, the new data can be uploaded to the site along with a unified metadata entry and an accompanying tutorial that improves reproducibility of the work and facilitates benchmarking by the community.

DATA SETS AND TUTORIALS

The original raw data for proteomics data sets currently included in ProteomicsML have already been made publicly available through ProteomeXchange,[15] mostly via the PRIDE database.[14] Here, the data hosted at ProteomicsML are provided in an ML-ready format, with links to original metadata and raw files for full provenance. Even though the data sets at ProteomicsML do not contain raw files, we do provide users with extensive tutorials on how to process raw data into ML-ready formats. ProteomicsML currently contains data sets and tutorials for fragmentation intensity, ion mobility (IM), retention time, and protein detectability. More data types can easily be added in the future, as the platform evolves along with the field.

(1) Retention time. Due to retention time playing a major role in modern peptide identification workflows, it is one of the most explored properties in predictive proteomics.[4] While some data sets for predicting retention time already exists, such as the publicly available data set from Kaggle (https://www.kaggle.com/datasets/kirillpe/proteomics-retention-time-prediction) and the DLOmix data sets (https://github.com/wilhelm-lab/dlomix/), we have also compiled new multitiered ML-ready data sets from the ProteomeTools synthetic peptide library,[13] in three specific sizes: 100 000 data points (small), well suited for new practitioners; (ii) 250 000 data points (medium), and (iii) 1 million data points (large), well suited for larger-scale ML training or benchmarking. As amino acid modifications can complicate the application of ML in proteomics, these three tiers do not contain any modified peptides except for carbamidomethylation of cysteine. Nevertheless, to train models for more real-life applications, we have also included an additional data set tier containing 200 000 oxidized peptides, as well as a

https://doi.org/10.1021/acs.jproteome.2c00629
J. Proteome Res. 2023, 22, 632−636

mixed data set containing 200 000 oxidized and 200 000 unmodified peptides. These data sets require minimal data preparation, although we still provide two distinct tutorials on methods to incorporate these data sets into deep learning (DL)-based models. In addition to preprocessed data, we also provide a detailed tutorial that combines and aligns retention times between runs from MaxQuant evidence files.[16] The output of this tutorial is a fully ML-ready file for retention time prediction.

(2) Fragmentation intensity. While it is easy to calculate the m/z values of theoretical peptide spectra, fragment ion peak intensities follow complex patterns that can be hard to predict. Nevertheless, these intensities can play a key role in accurate peptide identification.[17] For this reason, fragment ion intensity prediction is likely the second most explored topic for prediction purposes, for which comprehensive data sets and tutorials exist within ProteomicsML. As there are many attributes of peptides that affect their fragmentation patterns, the preprocessing steps of fragmentation data are more complex, and can be substantially different from lab to lab. For this reason, we have composed two separate tutorials, one that mimics the Prosit[6] data processing approach on the ProteomeTools[13] data sets, which consists of 745 000 annotated spectra, and one that mimics the MS^{2}PIP data process on a consensus human spectral library from the National Institute of Standards and Technology, which consists of 270 440 annotated spectra.[18] For data sets in this category it is difficult to provide a simple format with unified columns, as the handling and preprocessing steps differ significantly from model to model. Currently, there is one tutorial available on ProteomicsML describing the data processing pipeline from raw file to Prosit-style annotation, and we believe that with future additions we can provide users with tutorials for additional processing approaches.

(3) Ion mobility. Ion mobility is a technique to separate ionized analytes based on their size, shape, and physicochemical properties.[19] Techniques for ion mobility are generally based on propelling or trapping ions with an electric field in an ion mobility cell. Peptides are then separated by colliding them with an inert gas without fragmentation. Indeed, peptides with a larger area to collide will be more affected by the collisions, resulting in a higher measured collisional cross section (CCS). Historically, most methods predicting ion mobility were based on molecular dynamics models that calculate the CCS from first-principles in physics.[20] Lately the field has generated multiple ML and DL approaches for both peptide and metabolite CCS prediction.[21−23] The tutorials made available in ProteomicsML use both trapping (trapped ion mobility,[24] TIMS) and propelling ion mobility (traveling wave ion mobility,[25] TWIMS) data, where the large TIMS data set was sourced from Meier et al.[23] (718 917 data points) and the TWIMS data was sourced from Puyvelde et al.[26] (6268 data points). The tutorial is a walkthrough for training various model types, ranging from simple linear models to more complex nonlinear models (e.g., DL-based networks) showing advantages and disadvantages of various learning algorithms for CCS prediction.

(4) Protein detectability. Modern proteomics methods and instrumentation are now routinely detecting and quantifying the majority of proteins thought to be encoded by the genome of a given species.[27] Yet even after gathering enormous amounts of data, there is always a subset of proteins that remains refractory to detection. For example, even though tremendous effort has been focused on the human proteome, the fraction of unobserved proteins has been pushed just below 10%.[28,29] It remains unclear why certain proteins remain undetected, although ML has been applied to explore which properties most strongly influence detectability (as reviewed within).[30] One can compute a set of properties for a proteome and then train a model using those properties based on real world observations of the proteins that are detected and the proteins that are not detected. The model can be trained to learn which properties separate the detected from the undetected. Such a model has further utility to highlight proteins with properties that should sort them into the detected group, yet are not, as well as proteins that should belong to the undetected group, and yet they are detected. To facilitate this we have included the *Arabidopsis* PeptideAtlas data set (http://www.peptideatlas.org/builds/arabidopsis/), which is based on an extensive study of a single proteome.[31] This data set is based on the 2021 build, which has 52 data sets reprocessed to yield 40 million peptide-spectrum matches and a good overall coverage of the *Arabidopsis thaliana* proteome. Proteins in the data set are categorized as either "canonical", having the strongest evidence of detection, or "not observed", for which no peptides are identified. Along with these class labels, the data set contains various protein properties such as molecular weight, hydrophobicity, and isoelectric point, which could be crucial for classification purposes. The data set has an accompanying tutorial that illustrates how to analyze the data with a classification model for the observability of peptides.

Overall, these initial data set submissions and tutorials leave room for future expansion, until the community resource contains data sets for all properties previously and currently being explored in the field of proteomics. It is also open for user submissions, allowing researchers to upload their data in a standardized fashion, along with in-depth tutorials on their data handling and ML methodologies, resulting in more reproducible science. Our expectation is that this will shape the future of predictive proteomics, in favor of being more accessible, standardized, and reproducible.

Additionally, we have compiled a list of proteomics publications that utilize ML, along with a list of ProteomeXchange data sets used by each of the publications (Supplementary Table 1). Each of these ProteomeXchange data sets have been given a set of tags to indicate the nature of the usage in the publications (e.g., benchmarking, retention time, deep learning, etc.) as shown in Supplementary Table 2 (https://github.com/PRIDE-Utilities/pride-ontology/blob/master/pride-annotations/projects-proteomicsML.csv). Furthermore, these tags have also been added to the respective PRIDE data sets, which allows the tags to be easily searched, and for users to compile their ideal data set, if ProteomicsML does not already contain one.

https://doi.org/10.1021/acs.jproteome.2c00629
J. Proteome Res. 2023, 22, 632−636

CONCLUSION

We have presented ProteomicsML, a comprehensive resource of data sets and tutorials for every ML practitioner in the field of MS-based proteomics. ProteomicsML contains multiple data sets on a range of LC-MS peptide properties, allowing computational proteomics researchers to compare new algorithms to state-of-the-art models, as well as providing newcomers to the field with an accessible starting point, without requiring immediate in-depth knowledge of the entire proteomics analysis pipeline. We believe that this resource will aid the next generation of ML practitioners, and provide a stepping stone for more open and more reproducible science in the field.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00629.

Supplementary Table 1: Proteomics ML publications along with links to the ProteomeXchange data sets used for training or testing (XLSX)

Supplementary Table 2: Public ProteomeXchange data sets that have been used for ML training or benchmarking (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Juan Antonio Vizcaíno − *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, United Kingdom;* orcid.org/0000-0002-3905-4335; Phone: +44 (0) 1223 492686; Email: juan@ebi.ac.uk

Eric W. Deutsch − *Institute for Systems Biology, Seattle, Washington 98109, United States;* orcid.org/0000-0001-8732-0928; Phone: 206-732-1200; Email: edeutsch@systemsbiology.org; Fax: 206-732-1299

Authors

Tobias G. Rehfeldt − *Institute for Mathematics and Computer Science, University of Southern Denmark, 5000 Odense, Denmark;* orcid.org/0000-0002-1190-9485

Ralf Gabriels − *VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0002-1679-1711

Robbin Bouwmeester − *VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0001-6807-7029

Siegfried Gessulat − *MSAID Gmbh, Berlin 10559, Germany*

Benjamin A. Neely − *National Institute of Standards and Technology, Charleston, South Carolina 29412, United States;* orcid.org/0000-0001-6120-7695

Magnus Palmblad − *Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands;* orcid.org/0000-0002-5865-8994

Yasset Perez-Riverol − *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, United Kingdom;* orcid.org/0000-0001-6579-6941

Tobias Schmidt − *MSAID GmbH, Garching b. Munich 85748, Germany;* orcid.org/0000-0002-1883-6514

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00629

Author Contributions

[+]T.G.R., R.G., and R.B. contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): Tobias Schmidt and Siegfried Gessulat are employees of MSAID. MSAID makes ML-based software modules that are sold as part of Proteome Discoverer and also offers contract research. All other authors declare no competing financial interest.

Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor does it imply that the products identified are necessarily the best available for the proposed purpose.

ACKNOWLEDGMENTS

We thank Wassim Gabriel and Mathias Wilhelm for consultations on the Prosit annotation pipeline. The 2022 Lorentz Center workshop on Proteomics and Machine Learning was funded by the Dutch Research Council (NWO) with generous support from the Leiden University Medical Center, Thermo Fisher Scientific and *Journal of Proteome Research* (ACS). We also thank the staff at the Lorentz Center for helping make the hybrid workshop a success in pandemic times. T.G.R. acknowledges funding from the Velux Foundation [00028116]. R.G. acknowledges funding from the Research Foundation Flanders (FWO) [12B7123N]. R.B. acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen [HBC.2020.2205]. J.A.V. acknowledges funding from EMBL core funding, Wellcome [grant 223745/Z/21/Z], EU H2020 [823839], and BBSRC [BB/S01781X/1; BB/V018779/1]. E.W.D. acknowledges funding from the National Institutes of Health [R01 GM087221; R24 GM127667; U19 AG023122], and from the National Science Foundation [DBI-1933311; IOS-1922871].

REFERENCES

(1) von Heijne, G. Patterns of Amino Acids near Signal-Sequence Cleavage Sites. *Eur. J. Biochem.* **1983**, *133* (1), 17−21.

(2) Nielsen, H.; Brunak, S.; von Heijne, G. Machine Learning Approaches for the Prediction of Signal Peptides and Other Protein Sorting Signals. *Protein Eng.* **1999**, *12* (1), 3−9.

(3) Neely, B.; Dorfer, V.; Martens, L.; Bludau, I.; Bouwmeester, R.; Degroeve, S.; Deutsch, E. W.; Gessulat, S.; Käll, L.; Palczynski, P.; Payne, S.; Rehfeldt, T.; Schmidt, T.; Schwämmle, V.; Uszkoreit, J.; Vizcaíno, J. A.; Wilhelm, M.; Palmblad, M. Perspectives on Proteomics and Machine Learning. *J. Proteome Res.* **2023**, submitted.

(4) Wen, B.; Zeng, W.-F.; Liao, Y.; Shi, Z.; Savage, S. R.; Jiang, W.; Zhang, B. Deep Learning in Proteomics. *Proteomics* **2020**, *20* (21−22), No. e1900335.

(5) Bouwmeester, R.; Gabriels, R.; Van Den Bossche, T.; Martens, L.; Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* **2020**, *20* (21−22), No. e1900351.

(6) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M. Prosit:

Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nat. Methods* **2019**, *16* (6), 509−518.
(7) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry as-yet Unseen Modifications. *Nat. Methods* **2021**, *18* (11), 1363−1369.
(8) Meyer, J. G. Deep Learning Neural Network Tools for Proteomics. *Cell Rep. Methods* **2021**, *1* (2), No. 100003.
(9) Deng, L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine.* **2012**, *29*, 141−142.
(10) Fondrie, W. E.; Bittremieux, W.; Noble, W. S. Ppx: Programmatic Access to Proteomics Data Repositories. *J. Proteome Res.* **2021**, *20* (9), 4621−4624.
(11) Rehfeldt, T. G.; Krawczyk, K.; Bøgebjerg, M.; Schwämmle, V.; Röttger, R. MS2AI: Automated Repurposing of Public Peptide LC-MS Data for Machine Learning Applications. *Bioinformatics* **2022**, *38* (3), 875−877.
(12) Find Open Datasets and machine learning Projects. https://www.kaggle.com/datasets?search=proteomics (accessed 2022-10-02).
(13) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; Yu, P.; Schlegl, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building ProteomeTools Based on a Complete Synthetic Human Proteome. *Nat. Methods* **2017**, *14* (3), 259−262.
(14) 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. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543−D552.
(15) 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.; Ishihama, Y.; Vizcaíno, J. A. The ProteomeXchange Consortium in 2020: Enabling "Big Data" Approaches in Proteomics. *Nucleic Acids Res.* **2019**, *48* (D1), D1145−D1152.
(16) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11* (12), 2301−2319.
(17) Silva, A. S. C.; Bouwmeester, R.; Martens, L.; Degroeve, S. Accurate Peptide Fragmentation Predictions Allow Data Driven Approaches to Replace and Improve upon Proteomics Search Engine Scoring Functions. *Bioinformatics* **2019**, *35* (24), 5243−5248.
(18) Gabriels, R.; Martens, L.; Degroeve, S. Updated MS^2PIP Web Server Delivers Fast and Accurate MS^2 Peak Intensity Prediction for Multiple Fragmentation Methods, Instruments and Labeling Techniques. *Nucleic Acids Res.* **2019**, *47* (W1), W295−W299.
(19) Dodds, J. N.; Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (11), 2185−2195.
(20) Larriba-Andaluz, C.; Prell, J. S. Fundamentals of Ion Mobility in the Free Molecular Regime. Interlacing the Past, Present and Future of Ion Mobility Calculations. *Int. Rev. Phys. Chem.* **2020**, *39* (4), 569−623.
(21) Zhou, Z.; Xiong, X.; Zhu, Z.-J. MetCCS Predictor: A Web Server for Predicting Collision Cross-Section Values of Metabolites in Ion Mobility-Mass Spectrometry Based Metabolomics. *Bioinformatics* **2017**, *33* (14), 2235−2237.
(22) Broeckling, C. D.; Yao, L.; Isaac, G.; Gioioso, M.; Ianchis, V.; Vissers, J. P. C. Application of Predicted Collisional Cross Section to Metabolome Databases to Probabilistically Describe the Current and Future Ion Mobility Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (3), 661−669.
(23) Meier, F.; Köhler, N. D.; Brunner, A.-D.; Wanka, J.-M. H.; Voytik, E.; Strauss, M. T.; Theis, F. J.; Mann, M. Deep Learning the Collisional Cross Sections of the Peptide Universe from a Million Experimental Values. *Nat. Commun.* **2021**, *12* (1), 1185.
(24) Michelmann, K.; Silveira, J. A.; Ridgeway, M. E.; Park, M. A. Fundamentals of Trapped Ion Mobility Spectrometry. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (1), 14−24.
(25) Shvartsburg, A. A.; Smith, R. D. Fundamentals of Traveling Wave Ion Mobility Spectrometry. *Anal. Chem.* **2008**, *80* (24), 9689−9699.
(26) Van Puyvelde, B.; Daled, S.; Willems, S.; Gabriels, R.; Gonzalez de Peredo, A.; Chaoui, K.; Mouton-Barbosa, E.; Bouyssié, D.; Boonen, K.; Hughes, C. J.; Gethings, L. A.; Perez-Riverol, Y.; Bloomfield, N.; Tate, S.; Schiltz, O.; Martens, L.; Deforce, D.; Dhaenens, M. A Comprehensive LFQ Benchmark Dataset on Modern Day Acquisition Strategies in Proteomics. *Sci. Data* **2022**, *9* (1), 126.
(27) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. The One Hour Yeast Proteome. *Mol. Cell. Proteomics* **2014**, *13* (1), 339−347.
(28) Adhikari, S.; Nice, E. C.; Deutsch, E. W.; Lane, L.; Omenn, G. S.; Pennington, S. R.; Paik, Y.-K.; Overall, C. M.; Corrales, F. J.; Cristea, I. M.; Van Eyk, J. E.; Uhlén, M.; Lindskog, C.; Chan, D. W.; Bairoch, A.; Waddington, J. C.; Justice, J. L.; LaBaer, J.; Rodriguez, H.; He, F.; Kostrzewa, M.; Ping, P.; Gundry, R. L.; Stewart, P.; Srivastava, S.; Srivastava, S.; Nogueira, F. C. S.; Domont, G. B.; Vandenbrouck, Y.; Lam, M. P. Y.; Wennersten, S.; Vizcaino, J. A.; Wilkins, M.; Schwenk, J. M.; Lundberg, E.; Bandeira, N.; Marko-Varga, G.; Weintraub, S. T.; Pineau, C.; Kusebauch, U.; Moritz, R. L.; Ahn, S. B.; Palmblad, M.; Snyder, M. P.; Aebersold, R.; Baker, M. S. A High-Stringency Blueprint of the Human Proteome. *Nat. Commun.* **2020**, *11* (1), 5301.
(29) Omenn, G. S.; Lane, L.; Overall, C. M.; Paik, Y.-K.; Cristea, I. M.; Corrales, F. J.; Lindskog, C.; Weintraub, S.; Roehrl, M. H. A.; Liu, S.; Bandeira, N.; Srivastava, S.; Chen, Y.-J.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2021**, *20* (12), 5227−5240.
(30) Dincer, A. B.; Lu, Y.; Schweppe, D. K.; Oh, S.; Noble, W. S. Reducing Peptide Sequence Bias in Quantitative Mass Spectrometry Data with Machine Learning. *J. Proteome Res.* **2022**, *21* (7), 1771−1782.
(31) van Wijk, K. J.; Leppert, T.; Sun, Q.; Boguraev, S. S.; Sun, Z.; Mendoza, L.; Deutsch, E. W. The Arabidopsis PeptideAtlas: Harnessing Worldwide Proteomics Data to Create a Comprehensive Community Proteomics Resource. *Plant Cell* **2021**, *33* (11), 3421−3453.

https://doi.org/10.1021/acs.jproteome.2c00629
J. Proteome Res. 2023, 22, 632−636