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

Tobias G. Rehfeldt[1,*], Ralf Gabriels[2,*], Robbin Bouwmeester[2,*], Siegfried Gessulat[3], Benjamin A. Neely[4], Magnus Palmblad[5], Yasset Perez-Riverol[6], Tobias Schmidt[7], Juan Antonio Vizcaíno[6,+], Eric W. Deutsch[8,+]

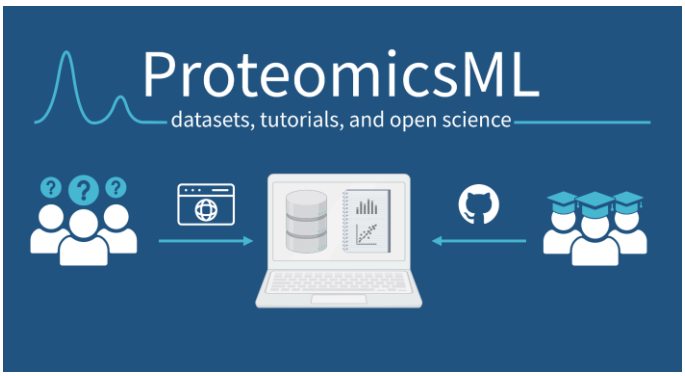
[1] Institute for Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[2] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium; Department of Biomolecular Medicine, Ghent University, Ghent, Belgium
[3] MSAID Gmbh, Berlin, Germany
[4] National Institute of Standards and Technology, Charleston, SC, USA
[5] Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands
[6] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
[7] MSAID GmbH, Garching b. Munich, Germany
[8] Institute for Systems Biology, Seattle WA 98109, USA

* These authors contributed equally to this work

Abstract

Dataset acquisition and curation are often the hardest and most time-consuming parts of a machine learning endeavor. This is especially true for proteomics-based LC-IM-MS datasets, due to the high-throughput data structure with high levels of noise and complexity between raw and machine learning-ready formats. While predictive proteomics is a field on the rise, when predicting peptide behavior in LC-IM-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 datasets 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 datasets that are useful for comparing state-of-the-art (SOTA) machine learning algorithms, as well as providing introductory material for teachers and newcomers to the field alike. The platform is freely available on https://www.proteomicsml.org/ and we welcome the entire proteomics community to contribute to the project at https://github.com/proteomicsml/.

Keywords: machine learning, deep learning, proteomics, educational platform, community platform, bioinformatics



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 range of physio-chemical properties associated with the field of high-throughput proteomics, as reviewed by Neely *et al.* (submitted, this issue). While many efforts are still in the realm of basic exploratory research, ML approaches are increasingly being incorporated into mainstream tools and standalone predictive resources.[3–6]

When training any ML model it is key to have suitable training and evaluation datasets. Likewise, in many fields of research where ML is applied, it is common to have a range of educational datasets, such as MNIST or IRIS, allowing newcomers to the field to easily learn common ML methodologies. Likewise, state-of-the-art (SOTA) models can use benchmarking datasets such as ImageNet or those available on the UCI Machine Learning Repository to compare their predictive capabilities. Similar to how the length of iris petals or the numbers of survivors of the Titanic have been modeled close to 50000 times[7], we seek to define proteomics datasets that can provide an entry point for ML modeling.

Although there have been numerous efforts to explore the predictive capabilities of models, there are several barriers that limit widespread adoption in the field of predictive proteomics. First, there are substantial difficulties in accessing datasets in a suitable form for ML applications. A substantial amount of effort is required to prepare raw proteomics datasets into a usable form, requiring expertise in proteomics data processing and intimate knowledge of the many post-processing methods available. Recently, tools such as ppx[8] and MS2AI[9] were created to facilitate this process, but they are still limited to certain use cases due to the complex nature of LC-IM-MS data.

Second, while some ML-ready datasets are available on platforms such as Kaggle[10] or in supplementary tables of publications, they are often difficult to find and lack long term maintenance and support post-publication. While there is no formal dataset consensus in the field, there are certain datasets that are often used for training such as ProteomeTools.[11] Nevertheless, there are no widely used datasets 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 post-processing protocols, such as normalization of measurements and re-scoring of PSMs.[3]

As an outcome of the 2022 Lorentz Center workshop on Proteomics and Machine Learning, 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 datasets that are ready to use for ML research, providing benchmark datasets for comparing different approaches, and encouraging new entrants into the field through expert-driven tutorials and other teaching materials.

Here we describe how the resource has been developed using commonly available tools and with future ease of maintenance in mind. We provide a brief overview of the datasets that are currently available at the resource and how it can be expanded 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 Resource

The primary entry point for the resource is the ProteomicsML.org website. It provides pages for general introductory datasets that are pre-processed and ready for training or evaluation, and pages for teaching resources and tutorials for those new to ML in proteomics. The code base for the website is maintained via a GitHub repository, and therefore is easy to maintain and amenable to outside

contributions from the field. We also collaborated with PRIDE to host larger datasets on a dedicated FTP server for ProteomicsML.

A key goal of ProteomicsML is to grow together with the field, which is why we provide experts with a contributing guide on how to upload datasets and tutorials for specific ML workflows or algorithms. After curation by the maintainers, contributions are automatically published on the website at ProteomicsML.org and are freely accessible for other researchers.

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

Along with well-annotated datasets, the platform provides users with in-depth tutorials on how to download, import, handle, and train various ML models. Many of the LC-IM-MS data types require certain, sometimes complicated, preprocessing steps in order to be fully compatible with ML frameworks. For this reason, we believe it to be 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 dataset-specific, allowing new tutorial submissions to focus on the direct interactions with specific ML models or methodologies, or to focus on a certain aspect of data preprocessing.

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

Data types

The original raw data for proteomics datasets currently included in ProteomicsML.org have already been made publicly available through ProteomeXchange[12], mostly via the PRIDE Archive.[13] Instead, 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 datasets at ProteomicsML do not contain raw files, we do aim to provide users with extensive tutorials on how to process raw data into ML-ready formats.

ProteomicsML currently contains datasets and tutorials for fragmentation intensity, ion mobility, 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. This is why we have provided multiple retention time datasets from various sources. We have combined several previously released ML-ready datasets - such as the Sharma et al. dataset from Kaggle and the DLOmix dataset - with a newly compiled multi-tiered dataset from the ProteomeTools synthetic peptide library.[11] From the latter, we have generated datasets of three sizes: 100,000 data points (small), well suited for newcomers; (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. Nevertheless, to train models for more real-life applications, we have also included an additional dataset tier containing 150,000 oxidized peptides, as well as a mixed dataset containing 150,000 oxidized and 150,000 unmodified peptides. These datasets require minimal data preparation, although we still provide two distinct tutorials on methods to incorporate these datasets 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.[14] 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.[15] For this reason, fragment ion intensity prediction is likely the second most explored topic, and which is why we choose to implement comprehensive datasets and tutorials for this data type. Since there are many attributes of peptides that affect their fragmentation patterns, the pre-processing 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 data processing approach on the ProteomeTools datasets, and one that mimics the MS^{2}PIP data process on a consensus human HCD dataset.[16] For datasets in this category it is difficult to provide a simple format with unified columns, as the handling and pre-processing 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 physio-chemical properties.[17] Initially the techniques for ion mobility propelled the ions with an electric field through a cell with inert gas where the ions collide with the inert gas without fragmentation. Separation is then achieved by the ions traveling faster or slower in the electric field (i.e., based on their charge) through the collisions with the gas (i.e., based on shape and size). Traveling wave ion mobility (TWIMS) works on the same principle but pushes the ions forward through the ion mobility cell with a wave of electric field.[18] Trapped ion mobility (TIMS) reverses this operation by trapping the ions in an electric field and forcing them forward by collision with the gas.[19] From any of the different ion mobility techniques one can derive the collisional cross-section (CCS) in Ångströms squared with the use of calibration analytes that have a known CCS. Historically most methods were based on molecular dynamics models that calculate the CCS from first principles in physics.[20] Lately the field has published multiple ML and DL approaches for both peptide and metabolite CCS prediction.[21–23] The tutorials made available in ProteomicsML use both TIMS and TWIMS data, where the large TIMS data set is from Meiers *et al.*[23] (718,917 data points) and the TWIMS data is from Puyvelde *et al.*[24] (6268 data points). The tutorial is a walkthrough that trains linear models to more complex non-linear models (e.g., DL based networks) showing advantages and disadvantages of the learning algorithms for CCS prediction.

(4) Protein detectability. Modern proteomic methods and instrumentation are now routinely detecting and quantifying the majority of proteins thought to be encoded by the genome of a species.[25] Yet even after gathering enormous amounts of data, there is always a subset of proteins that remains refractory to detection. For example, even through tremendous effort focused on the human proteome, the fraction of unobserved proteins has been pushed just below 10%.[26,27] It remains unclear why certain proteins remain undetected, though machine learning has been applied to explore which properties most strongly influence detectability (as reviewed within).[28] 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 aren't 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 that have properties that should make them belong to the detected group, but 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 a dataset that is based on an extensive study of a proteome: the Arabidopsis PeptideAtlas.[29,30] This dataset is based on the 2021 build, which has 52 datasets reprocessed to yield 40 million peptide-spectrum matches and good coverage of the *Arabidopsis thaliana* proteome. Proteins in the dataset are categorized as either "canonical", the strongest evidence of detection, or "not observed" if known peptides are not identified. Along with these class labels, the dataset contains various protein properties such as molecular weight, hydrophobicity, and isoelectric point that could be crucial for classification purposes. The dataset has an accompanying tutorial that illustrates how to analyze the data with a multilayer perceptron model to classify the observability of peptides.

Overall, these initial dataset submissions and tutorials leave room for a range of future expansion, until the community resource contains datasets 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 for more reproducible science, along with in-depth tutorials on their data handling and ML methodologies. Our hope is that this will shape the future of predictive proteomics, in favor of being more introducible, standardized and reproducible.

Additionally, we have compiled a list of proteomics publications that utilize ML, along with a list of ProteomeXchange datasets used by each of the publications (Supplementary Table 1). Each of these ProteomeXchange datasets 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 seen in Supplementary Table 2[31]. Furthermore, these tags have also been added to the respective PRIDE entries, which allows the tags to easily be searched, and for users to compile their ideal dataset, if ProteomicsML does not already contain one.

Conclusion

We have presented ProteomicsML.org, a comprehensive resource of datasets and tutorials for every ML practitioner in the field of MS-based proteomics. ProteomicsML contains multiple datasets on a range of LC-IM-MS peptide properties, allowing computational proteomics researchers to compare new algorithms to the state-of-the-art models, as well as providing newcomers to the field with an easier 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 reproducible science in the field.

Supporting Information

Supplementary Table 1: Proteomics ML publications along with links to the ProteomeXchange datasets used for training or testing.

Supplementary Table 2: Public ProteomeXchange datasets that have been used for ML training or benchmarking.

Author Information

+Address correspondence to:
Eric Deutsch: Email: edeutsch@systemsbiology.org, Phone: 206-732-1200, Fax: 206-732-1299
Juan Antonio Vizcaíno: Email: juan@ebi.ac.uk,

Notes

Tobias Schmidt and Siegfried Gessulat are employees of MSAID. MSAID makes machine learning-based software modules that are sold as part of Proteome Discoverer and also offers contract research.

Acknowledgements

We would like to 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 the Journal of Proteome Research. Thanks also goes out to the staff at the Lorentz Center for helping make the hybrid workshop a success in pandemic times. Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose. R.B. acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen under project number HBC.2020.2205. R.G. acknowledges funding from the Research Foundation Flanders (FWO) [12B7123N]. T.G.R acknowledges funding from the Velux Foundation [00028116]. E.W.D. acknowledges funding from National Institutes of Health grants R01GM087221, R24GM127667, U19AG023122, and by National Science Foundation grants DBI-1933311, and IOS-1922871. JAV acknowledges funding from EMBL core funding, EU H2020 [grant number 823839], and BBSRC [grant numbers BB/S01781X/1 and BB/V018779/1].

References

(1) von Heijne, G. Patterns of Amino Acids near Signal-Sequence Cleavage Sites. *Eur. J. Biochem.* **1983**, *133* (1), 17–21. https://doi.org/10.1111/j.1432-1033.1983.tb07424.x.

(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. https://doi.org/10.1093/protein/12.1.3.

(3) 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. https://doi.org/10.1038/s41592-019-0426-7.

(4) 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. https://doi.org/10.1038/s41592-021-01301-5.

(5) 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), e1900335. https://doi.org/10.1002/pmic.201900335.

(6) Meyer, J. G. Deep Learning Neural Network Tools for Proteomics. *Cell Rep Methods* **2021**, *1* (2), 100003. https://doi.org/10.1016/j.crmeth.2021.100003.

(7) *Titanic - machine learning from disaster*. https://www.kaggle.com/competitions/titanic (accessed 2022-10-02).

(8) Fondrie, W. E.; Bittremieux, W.; Noble, W. S. Ppx: Programmatic Access to Proteomics Data Repositories. *J. Proteome Res.* **2021**, *20* (9), 4621–4624. https://doi.org/10.1021/acs.jproteome.1c00454.

(9) 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* **2021**. https://doi.org/10.1093/bioinformatics/btab701.

(10) *Find Open Datasets and machine learning Projects*. https://www.kaggle.com/datasets?search=proteomics (accessed 2022-10-02).

(11) 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. https://doi.org/10.1038/nmeth.4153.

(12) 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.* **2020**, *48* (D1), D1145–D1152. https://doi.org/10.1093/nar/gkz984.

(13) 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. https://doi.org/10.1093/nar/gkab1038.

(14) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11* (12), 2301–2319. https://doi.org/10.1038/nprot.2016.136.

(15) C Silva, A. S.; 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. https://doi.org/10.1093/bioinformatics/btz383.

(16) Gabriels, R.; Martens, L.; Degroeve, S. Updated MS^{2}PIP 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. https://doi.org/10.1093/nar/gkz299.

(17) 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. https://doi.org/10.1007/s13361-019-02288-2.

(18) Shvartsburg, A. A.; Smith, R. D. Fundamentals of Traveling Wave Ion Mobility Spectrometry. *Anal. Chem.* **2008**, *80* (24), 9689–9699. https://doi.org/10.1021/ac8016295.

(19) 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. https://doi.org/10.1007/s13361-014-0999-4.

(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.

https://doi.org/10.1080/0144235X.2020.1826708.
(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. https://doi.org/10.1093/bioinformatics/btx140.
(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. https://doi.org/10.1021/jasms.0c00375.
(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. https://doi.org/10.1038/s41467-021-21352-8.
(24) 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. https://doi.org/10.1038/s41597-022-01216-6.
(25) 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. https://doi.org/10.1074/mcp.M113.034769.
(26) 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. https://doi.org/10.1038/s41467-020-19045-9.
(27) 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. https://doi.org/10.1021/acs.jproteome.1c00590.
(28) 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. https://doi.org/10.1021/acs.jproteome.2c00211.
(29) 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/10.1093/plcell/koab211.
(30) *Arabidopsis PeptideAtlas*. http://www.peptideatlas.org/builds/arabidopsis/ (accessed 2022-09-30).
(31) *Projects-proteomicsML.csv at Master · PRIDE-Utilities/pride-Ontology*; Github.