Boosting the MS1-only proteomics with machine learning allows 2000 protein identifications in 5-minute proteome analysis

Mark V. Ivanov^{1*}, Julia A. Bubis¹, Vladimir Gorshkov², Daniil A. Abdrakhimov^{1,3}, Frank Kjeldsen², Mikhail V. Gorshkov¹

¹ V. L. Talrose Institute for Energy Problems of Chemical Physics, N. N. Semenov Federal Research Center for Chemical Physics, Russian Academy of Sciences, 38 Leninsky Pr., Bld. 2, Moscow, 119334, Russia

²Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense M, Denmark

³Moscow Institute of Physics and Technology, Institutsky lane 9, Dolgoprudny, Moscow region, 141700, Russia

ABSTRACT

Proteome-wide analyses most often rely on tandem mass spectrometry imposing considerable instrumental time consumption that is one of the main obstacles in a broader acceptance of proteomics in biomedical and clinical research. Recently, we presented a fast proteomic method termed DirectMS1 based on MS1-only mass spectra acquisition and data processing. The method allowed significant squeezing of the proteome-wide analysis to a few minute time frame at the depth of quantitative proteome coverage of 1000 proteins at 1% FDR. In this work, to further increase the capabilities of the DirectMS1 method, we explored the opportunities presented by the recent progress in the machine learning area and applied the LightGBM tree-based learning algorithm into the scoring of peptide-feature matches when processing MS1 spectra. Further, we integrated the peptide feature identification algorithm of DirectMS1 with the recently introduced peptide retention time prediction utility, DeepLC. Additional approaches to improve performance of the DirectMS1

method are discussed and demonstrated, such as FAIMS coupled to the Orbitrap mass analyzer. As a result of all improvements to DirectMS1, we succeeded in identifying more than 2000 proteins at 1% FDR from the HeLa cell line in a 5 minute LC-MS1 analysis.

Keywords: protein identification, mass spectrometry, high-throughput proteomics, machine learning, retention time prediction.

INTRODUCTION

High throughput and sensitive analytical approaches enabling proteome-wide measurements of large cohorts of samples will open the way for application of mass spectrometry-based proteomics in clinical trials, population proteomics, as well as in emerging areas of drug-to-proteome interactions and metaproteome characterization.^{1,2} Moreover, the needed throughput and depth of proteome coverage in these studies has to be accompanied with protein quantitation consistent across the sample cohorts, which is necessary for these approaches being useful in personalized medicine studies.^{3,4} A recent example on using the large cohort of COVID-19 patients clearly demonstrated the need for ultra-high-throughput proteomics to generate hypotheses about therapeutic targets and aid classification or diagnostic decision making in clinical environments.⁵

The typical MS/MS-based proteomic methods are extremely instrument time consuming, which is partially overcome by expensive multiplexing using isotopic labeling strategies, such as tandem mass tag (TMT).⁶ Nevertheless, a number of recent studies employing the bottom-up proteomic approaches become increasingly focused on increasing the throughput of the proteome-wide analysis. These studies include all steps of the proteome characterization workflow, such as increasing the speed of sample preparation prior to MS-based analysis⁷⁻¹¹, squeezing data

acquisition time to a few minute range¹²⁻¹⁴ and increasing the throughput of data processing¹⁵⁻²⁰.

One of the most time-consuming parts of the analysis is MS/MS data acquisition, which requires using long HPLC gradient times to have enough room for sequential isolation of as many as possible precursor ions from each MS1 spectrum for subsequent fragmentation. A number of methods and approaches to MS/MS-free proteome analysis have been proposed and widely explored starting from the early Accurate Mass and Tag (Time) method²¹⁻²⁵ to a recent truly (e.g., without employing tandem mass spectrometry at any step in the workflow) MS/MS-free realizations^{26,27}. These methods rely heavily on the accuracies of both peptide m/z measurements and retention time (RT) predictions. The latter is especially important for the MS/MS-free strategy as only retention times contain sequence-specific information in a (m/z , RT) space.^{28,29} With advances in development of machine learning algorithms a variety of highly accurate RT prediction models become increasingly available.³⁰⁻³³ The latter work is particularly interesting as it shows that DeepLC model's performance is comparable with the other deep learning-based alternatives, yet, it provides better generalization between different chromatography setups. This feature is particularly useful for MS/MS-free proteomic approaches such as DirectMS1, when the number of peptides available for RT prediction model training is typically small. Therefore, the use of a pre-trained DeepLC model for rapid adoption to a dataset obtained for particular separation conditions becomes advantageous.

Previously, we described the DirectMS1 method, in which proteolytic peptide mixture analysis is performed in MS/MS-free mode of acquisition using high resolution mass spectrometry.²⁷ Because the method does not employ isolation and fragmentation steps, the time for peptide separation can be reduced significantly to few minutes and the number of MS1 spectra available for processing will only be limited by the acquisition rate of the mass analyzer operating at high mass resolution. For the first time, DirectMS1 demonstrated the capability to identify up to 1000 proteins from a HeLa cell line using only a 5-minute HPLC gradient. Moreover, the average sequence coverage for each identified protein in this method exceeded the one of a standard MS/MS-based approach (even when long

gradient is used) by almost an order of magnitude, thus, significantly improving the quantitation. On a proteome-wide scale this kind of analysis efficiency was not considered feasible before because of the whole proteome sample complexity, a lack of sequence specific information in the measured m/z values of peptide ions, and the low accuracy of existing phenomenological retention time prediction models. However, advances in high resolution mass spectrometry technology and machine learning-based RT prediction models implemented in DirectMS1's search engine, allowed the method to exceed the proteome coverage rate per analysis time of a long gradient MS-MS-based proteomics. Yet, the method's efficiency does not reflect the real information content of the MS1 dataset, in which all eluted and ionizable peptides of the analyzed proteolytic mixture are potentially present in a full dynamic range accessible by current state-of-the-art high resolution mass analyzers.

In this work, we integrated two novel machine learning algorithms into the data processing workflow of DirectMS1 to further improve its efficiency. These algorithms include DeepLC as a retention time prediction model used in the method's peptide identification and the gradient boosted machine, LightGBM³⁴, for scoring peptide-feature matches. Also, the method was upgraded and tested for processing MS1-only data obtained using high resolution mass spectrometry and the high-field asymmetric waveform ion mobility, FAIMS³⁵, which was found increasingly applicable in proteomic research^{36,37} as it provides additional separation dimension for peptides at the front end of a mass spectrometer.

EXPERIMENTAL SECTION

Samples. All experiments were performed using Thermo Scientific Pierce™ HeLa Protein Digest Standard (P/N 88328) derived from HeLa S3 cell line.

Data acquisition. LC-MS analysis in this work was performed using Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific, San Jose, CA, USA) coupled with UltiMate 3000 LC system (Thermo Fisher Scientific, Germering, Germany) and equipped with *FAIMS Pro* interface. Short LC gradient method was applied as reported previously²⁷. Trap column μ -

Precolumn C18 PepMap100 (5 μ m, 300 μ m, i.d. 5 mm, 100 Å) (Thermo Fisher Scientific, USA) and self-packed analytical column (Inertsil 3 μ m, 75 μ m i.d., 15 cm length) were employed for separation. Mobile phases were as follows: (A) 0.1 % formic acid (FA) in water; (B) 80 % acetonitrile (ACN), 0.1 % FA in water. Loading solvent was 0.05 % trifluoroacetic acid (TFA) in water. The gradient was from 5 % to 35 % phase B in 4.8 min at 1.5 μ L/min. Total method time including column washing and equilibration was 7.3 min. Field asymmetric ion mobility spectrometry (FAIMS) separations were performed with the following settings: inner and outer electrode temperatures were 100 °C; FAIMS carrier gas flow was 4.7 L/min; asymmetric waveform dispersion voltage (DV) was –5000 V; entrance plate voltage was 250 V. Compensation voltages (CV) -50 V, -65 V, and -80 V were used in a stepwise mode during LC-MS analysis. Mass spectrometry measurements were performed in MS1-only mode of acquisition. Full MS scans were acquired in a range from m/z 375 to 1500 at a resolution of 120 000 at m/z 200 with Automatic Gain Control (AGC) target of $4 \cdot 10^5$, 1 microscan and 50 ms maximum injection time. Two hundred ng of HeLa digest was loaded on column.

Experimental data obtained previously for 500 ng HeLa injection in both MS1-only MS/MS using Q-Exactive HF mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) were taken from PRIDE repository³⁸ (repository identifier PXD015669) and used for testing upgraded DirectMS1 software. Supplementary Table S1 provides a list and the annotations for all experimental data used in this work.

Data Analysis. Raw files were converted into mzML format using *msConvert* from ProteoWizard (v. 3.0.20066).³⁹ Peptide features were detected using either Dinosaur⁴⁰ (v. 1.2.0) or in-house developed Biosaur (freely available at <https://github.com/abdrakhimov1/Biosaur>) (v. 1.0.3) software for data obtained without and with FAIMS, respectively. Proteins were identified by using *ms1searchpy* (v. 2.0.3), which is the search engine of DirectMS1. This software is open-source and freely available at <https://github.com/markmpt/ms1searchpy> under Apache 2.0 license. Parameters for the search were as follows: minimum 3 scans for detected peptide isotopic cluster; minimum one visible C13 isotope; charges from 1+ to 6+, no missed cleavage sites, and 8 ppm initial mass accuracy. All

searches were performed against Swiss-Prot human concatenated database containing 20247 protein sequences and its decoys unless otherwise stated. Results were filtered to 1% protein level false discovery rate (FDR) using target-decoy approach⁴¹ with its “picked” modification⁴² and “+1” correction⁴³.

Data Availability. The datasets generated and analyzed during the current study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD022094.

RESULTS AND DISCUSSION

Figure 1. DirectMS1 workflow.

The workflow. Figure 1 shows details of DirectMS1 method. The method starts with acquiring high resolution peptide ion mass spectra. A mass spectrometer operates in MS1-only mode and simply collects spectra for eluting peptides at the speed determined by the AGC and mass resolution

settings. The total number of MS1 spectra acquired during 5-min LC gradient at the mass resolution of 120,000 at m/z 200 and AGC of 3×10^6 ranges from 1,000 to 1,500 depending on the mass spectrometer model.

Second step is peptide features detection using either Dinosaur or Biosaur utilities. The latter was developed in-house recently and allows handling ion mobility data, such as FAIMS. The third step is preliminary search performed using only the information about measured peptide m/z values. High confident peptide feature matches (PFMs), the ones belonging to the top scored proteins identified in the preliminary search and having the smallest mass error, are selected for RT prediction models calibration in the next step. By default, the first 2500 PFMs with the highest confidence are used, many of them being false matches (see discussion below).

Training of peptide RT prediction models integrated into the search workflow is one of the crucial steps in DirectMS1. Originally it was having an option to use one of the two RT models. The first one was relying on the so-called “additive” retention time prediction model, in which a peptide retention time is a sum of individual retention coefficients of residues constituting the peptide sequence.⁴⁴ These coefficients can be either predefined for the type of used chromatography, or calculated on-the-fly for specific data and separation conditions. The second model was based on the ELUDE³⁰ RT prediction software and in general allows significantly more accurate prediction of peptide retention times compared with the additive model, although for the price of longer computation time. In this work we upgraded DirectMS1 search module by integrating it with the recently introduced deep learning RT prediction model - DeepLC³³. It performs reportedly as well, or better, as the other state-of-the-art retention time prediction models employing machine learning algorithms. Indeed, DeepLC model allowed significant improvement in the accuracy of retention time prediction for identifying peptides in DirectMS1 compared with the previously used additive and ELUDE models that resulted in dramatic increase in the number of identified proteins as shown in Figure 2a and b. Furthermore, a two-stage calibration strategy was implemented, which allowed to further decrease RT prediction error and increase the

number of identifications as detailed below in the *Retention time prediction models* section.

At the fifth step of the workflow, a gradient boosting algorithm is applied for estimating the PFM's quality. The final search uses PFM's quality information for protein score calculations and prepares protein identification list at the specified FDR level, typically 1% (see discussion below).

Figure 2. *Left column.* Analysis of HeLa cell line runs using DirectMS1: **(a)** Number of protein groups identified at 1% FDR for different RT training strategies. “One-stage” denotes standard RT training, “two-stage” -- RT training using two-stage strategy with the same prediction model for both stages, and “two-stage*” -- RT training using additive and DeepLC models at first and second stages of training, respectively; **(b)** Retention time prediction accuracy determined as a standard deviation in the distribution of peptide matches of identified proteins by the difference between experimental and predicted RT for different RT training strategies. *Middle column.* performance of DirectMS1 with different initial training sets for RT prediction: **(c)** Identified protein groups at 1% FDR; and **(d)** Standard deviation of difference between experimental and predicted retention time for reliable peptide matches. The results are shown for DirectMS1 analysis of HeLa cell line using LC-MS1-only and DDA data with excluded MS/MS spectra. Green color is DDA MS/MS analysis results. *Right column.* Number of identified protein groups at 1% FDR for DirectMS1 with gradient

boosting model, LightGBM, turned on/off: **(e)** Q-Exactive HF; and **(f)** Fusion Lumos with FAIMS setup. Results are shown in average for 3 replicates, as well as for all replicates combined into single identification results.

Retention time prediction models. Retention time prediction procedure in DirectMS1 has an important distinction compared with the typically used in conventional MS/MS-based searches. The latter usually uses a list of high-confident peptide-spectrum matches (PSM) (typically filtered to 1% FDR level) for RT model training, retraining, or calibration. On the contrary, identifications in DirectMS1 have a high rate of false positives among PFMs and it is not possible to obtain a subset of high-confident peptides of the size reasonable for RT training. DirectMS1 begins with a preliminary search as shown in Fig.1 as the second step of the workflow using only m/z values of detected PFMs. Then, it forms a preliminary list of identified proteins at 1% FDR. Each of these proteins contains a set of PFMs, with some of them being wrong, and these PFMs have different levels of deviations of experimentally determined m/z from the theoretically calculated one. PFMs with the lowest mass deviations, up to 2500 PFMs in case of the cellular proteome analyses, are further taken for training the DeepLC model. Also note that DirectMS1 uses RT prediction models in different ways. While the additive model is trained from scratch with no use of previously trained models, both ELUDE and DeepLC are the pretrained models re-calibrated for the current data. Importantly, the results of these recalibrations can be affected by the presence of false positives matches in the training set of PFMs. To alleviate this problem, DirectMS1 employs two-stage RT calibration as shown in Figure 1, bottom panel. First, the model is trained using all 2500 peptides from the preliminary training set. Then, the software calculates the differences between predicted and experimental retention times and fits this difference array with the sum of Gaussian and uniform distributions. Here, we assume that true peptide matches should follow the Gaussian distribution, while the false matches follow the uniform one. Using the standard deviation obtained for the distribution of all PFMs by differences between experimental and predicted RTs, peptide matches with RT differences exceeding three standard deviations are excluded from the

training set. Then, the model is subjected to final training using the updated PFM set. Figure 2a and b illustrate how different RT prediction models and one- or two-stage training procedures affect the performance of DirectMS1 method. The results show that the additive model benefits the most from the two-stage RT training process, although it still underperforms compared with machine learning based approaches, which are probably less influenced by the presence of false matches in the training set. However, the two-stage RT training where the first stage is performed using the additive model followed by training DeepLC by the set of RT filtered PFMs as described above provides the maximal number of identified proteins. This two-stage training involving two different RT prediction models is denoted as *two-stage** in Figure 2 and is currently used by default in DirectMS1's search engine. All further results presented here for demonstration of the efficiency of the method have been obtained using the *two-stage** training.

Peptide set for RT prediction models. DirectMS1 identifies proteins without use of any MS/MS spectra. However, one can expect that having MS/MS-based identifications for RT model training may potentially benefit the DirectMS1 analysis. This training could be done in two ways: (i) all data are acquired in DDA mode and MS1 spectra are then processed by DirectMS1, while MS/MS identifications are used for RT model calibration; and (ii) a separate single DDA run is performed, which is used for RT model training for several subsequent MS1-only experiments. Both approaches have some drawbacks. Separate DDA run require additional experimental time and sample and can also result in RT alignment errors.⁴⁵ On the other hand, analyzing data using standard DDA mode reduces the number of available MS1 spectra, reducing DirectMS1 performance, or requires longer gradient times for the whole proteome characterization undermining the concept of high throughput proteomics. Also, both above-mentioned approaches for RT model calibration require a mass spectrometer capable for MS/MS mode, while DirectMS1 method can be used with simpler and more robust MS1-only instrumentation. To further elaborate on the issue we calculated the number of identified protein groups at 1% FDR and RT prediction error obtained using DeepLC model

for different approaches as shown in Figure 2c and d. Firstly, DirectMS1 applied to DDA data shows that RT calibration using peptides identified in MS1 analysis is as good as the one based on the peptides identified from MS/MS spectra. Secondly, DirectMS1 applied to MS1-only data shows that transferring RT calibration results between datasets decreases the overall RT prediction accuracy.

Machine learning analysis of PFMs. Distributions of PFMs' mass and RT errors are shown in Figure 3 a, b for target and decoy matches. Expectedly, the mass error does not provide a strong differentiation between target and decoy PFMs as shown in Fig 3a. On the other hand, retention time difference between experimental and predicted values present potentially a strong discrimination metric since peptide retention times are sequence-specific (Figure 3b). Note also, that there can be some other peptide properties discriminating targets and decoys. For example, the RT prediction and mass measurement errors can depend on peptide lengths and signal intensities, respectively. Gradient boosting techniques should potentially account for these non-linear effects and use them to increase the discriminative power of the search engine. In DirectMS1's search engine *ms1searchpy* we employed the LightGBM gradient boosting framework for estimating the PFM quality score. A list of target-decoy discriminative properties used for LightGBM in the current version of *ms1searchpy* is shown in Supplementary Table S2. Hyperparameters for LightGBM are tuned in DirectMS1 on-the-fly during data processing. It is implemented using *Random Search Cross-Validation method* with 25 random hyperparameters testing and 3-fold group based cross-validation. Unique peptide sequences are chosen as groups to prevent overfitting by the presence of the same sequences in both training and test cross-validation sets. The optimal hyperparameters which provide the maximal average number of PFMs at 25% FDR in the test groups are chosen for final LightGBM model training. Calculated values from gradient boosting are further used in the iteration procedure for protein score calculation. Firstly, 10% of PFMs with best LightGBM scores are picked and protein binomial scores are calculated using only these PFMs. At the next iteration, 20% of PFMs are picked and scores are calculated again. It continues with

10% step up to 100%, and in the end 10 intermediate protein scores become available for each protein at the end. These scores are averaged and used as the final protein score for target-decoy filtering and generating the list with protein identification results.

Comparison of DirectMS1 performance with and without the LightGBM framework is shown in Figure 2e. Using this framework adds 7 to 10% of protein identifications. Combining multiple replicates into the single output increases further the number of protein identifications at 1% FDR to nearly 1900 for this particular dataset.

Figure 3. DirectMS1 analysis for 5 min LC-MS data of 200 ng of HeLa lysate using Lumos with FAIMS interface: **(a)** distributions of target and decoy PFMs for mass difference between theoretical and experimental values; **(b)** distributions of target and decoy PFMs for retention time difference between experimental and predicted values; **(c)** estimated peptide-level FDR for a cumulative set of top scored proteins; **(d)** percentage of *E. Coli* proteins in the results at 1% FDR. Here, light blue color is the results for single replicates averaged over 3 technical replicates, and dark blue color is the results for a combination of 3 technical replicates before protein score calculation. Dashed black line is the expected percentage of *E. Coli* proteins in the results which is a fraction of *E. Coli* proteins in the search database.

FAIMS data. FAIMS is an atmospheric pressure ion mobility technique that separates gas-phase ions by the difference in their mobility in high and low electric fields. FAIMS separation is orthogonal³⁶ to both LC and MS and provides additional on-line fractionation to further improve dynamic range of the analysis of complex samples. Having this third separation dimension in addition to masses and retention times can be beneficial for

DirectMS1 method despite the decrease, albeit marginal, in the average number of MS1 scans per peptide feature when multiple compensation voltages are used. Although the prediction models for FAIMS data are yet to be developed, adding this dimension into the DirectMS1 workflow resulted in the meaningful increase in the number of identified proteins. As shown in Fig.2f, we were able to identify 2100 proteins at 1% FDR in 5-min HeLa proteome analysis after combining 3 technical replicates.

Validation. False discovery rate at peptide level can be estimated by the following formula:

$$FDR_{peptide} = p \times \frac{\text{number of theoretical peptides}}{\text{number of matched peptides}} \quad (1),$$

where

$$p = \frac{\text{number of matched decoy peptides}}{\text{number of theoretical decoy peptides}} \quad (2).$$

False discovery rate analysis at peptide level is shown in Figure 3c. Strikingly enough, the average FDR for all peptide-feature matches is 90 %. However, the average peptide-level FDR values are 23 % and 32 % for top 100 scored proteins and all proteins at 1% protein FDR level, respectively.

While protein FDR level is controlled by target-decoy approach, the other interesting question is how many proteins from the target database not present in the sample can be found in the results by DirectMS1's search engine? To perform this control analysis, HeLa data was searched against combined human and *E. coli* Swiss-Prot databases. The number of identified *E. coli* proteins in DirectMS1 analysis is shown in Fig.3d. The expected percentage of *E. coli* proteins is calculated by the formula:

$$\text{Expected percentage of } E. coli \text{ proteins} = FDR \times \frac{\text{Number of } E. coli \text{ proteins in database}}{\text{Total number of proteins in database}} \quad (3)$$

from which we found that 0.18% of *E. coli* proteins at 1% FDR in the identification results as expected (*E. coli* and human databases contain 8862 and 40386 sequences including decoys, respectively). This result

shows that DirectMS1 provides an accurate threshold for filtering protein identifications.

CONCLUSIONS

DirectMS1 for high-throughput MS/MS-free quantitative proteome analysis was upgraded with DeepLC, the novel machine learning-based utility for peptide retention time prediction and the recently introduced gradient boosting framework LightGBM for scoring peptide feature matches in MS1 spectra. Additionally, two-stage and two-model peptide retention time prediction was included into the method's workflow. Finally, DirectMS1 was tested using a high-resolution mass spectrometer with FAIMS interface that brought the third, peptide structure-dependent dimension into the MS1-only search space. These additions resulted in a two-fold increase in DirectMS1 method's efficiency which is now capable of identifying as much as 2000 proteins at 1% FDR in a 5-min MS1-only analysis of cellular proteomes. Further improvement in the depth of proteome coverage using this method are expected with improvements in ion mobility data processing and prediction.

ASSOCIATED CONTENT

Supporting information:

Supplementary Table S1. Raw files description with experimental parameters.

Supplementary Table S2. A list of target-decoy discriminative properties used for LightGBM.

ACKNOWLEDGMENTS

The authors thank Prof. Lukas Kall and Dr. Matthew The for helpful discussion and suggestions regarding machine learning analysis of PFMs. Proteomic data processing, statistical analysis, and software development were performed with financial support from Russian Science Foundation, Grant no. 20-14-00229 to M.V.G. Proteomics and mass spectrometry research at University of Southern Denmark SDU are supported by generous grants to the VILLUM Center for Bioanalytical Sciences (VILLUM Foundation grant no. 7292) and PRO-MS: Danish National Mass Spectrometry Platform for Functional Proteomics (grant no. 5072-00007B).

REFERENCES

- (1) Bache, N.; Geyer, P. E.; Bekker-Jensen, D. B.; Hoerning, O.; Falkenby, L.; Treit, P. V.; Doll, S.; Paron, I.; Müller, J. B.; Meier, F.; Olsen, J. V.; Vorm, O.; Mann, M. A Novel LC System Embeds Analytes in Pre-Formed Gradients for Rapid, Ultra-Robust Proteomics. *Mol. Cell Proteomics* **2018**, 17 (11), 2284-2296. <https://doi.org/10.1074/mcp.TIR118.000853>.
- (2) Baker, E. S.; Liu, T.; Petyuk, V. A.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Anderson, G. A.; Smith, R. D. Mass Spectrometry for Translational Proteomics: Progress and Clinical Implications. *Genome Med* **2012**, 4 (8), 63. <https://doi.org/10.1186/gm364>.
- (3) Gillet, L. C.; Leitner, A.; Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**, 9 (1), 449-472. <https://doi.org/10.1146/annurev-anchem-071015-041535>.
- (4) Zecha, J.; Lee, C.-Y.; Bayer, F. P.; Meng, C.; Grass, V.; Zerweck, J.; Schnatbaum, K.; Michler, T.; Pichlmair, A.; Ludwig, C.; Kuster, B. Data, Reagents, Assays and Merits of Proteomics for SARS-CoV-2 Research and Testing. *Mol Cell Proteomics* **2020**, 19 (9), 1503-1522. <https://doi.org/10.1074/mcp.RA120.002164>.
- (5) Messner, C. B.; Demichev, V.; Wendisch, D.; Michalick, L.; White, M.; Freiwald, A.; Textoris-Taube, K.; Vernardis, S. I.; Egger, A.-S.; Kreidl,

- M.; Ludwig, D.; Kilian, C.; Agostini, F.; Zelezniak, A.; Thibeault, C.; Pfeiffer, M.; Hippenstiel, S.; Hocke, A.; von Kalle, C.; Campbell, A.; Hayward, C.; Porteous, D. J.; Marioni, R. E.; Langenberg, C.; Lilley, K. S.; Kuebler, W. M.; Mülleder, M.; Drosten, C.; Suttorp, N.; Witzernath, M.; Kurth, F.; Sander, L. E.; Ralser, M. Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Syst* **2020**, *11* (1), 11-24.e4. <https://doi.org/10.1016/j.cels.2020.05.012>.
- (6) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K. A.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal Chem* **2003**, *75* (8), 1895-1904. <https://doi.org/10.1021/ac0262560>.
- (7) Tubaon, R. M.; Haddad, P. R.; Quirino, J. P. Sample Clean-up Strategies for ESI Mass Spectrometry Applications in Bottom-up Proteomics: Trends from 2012 to 2016. *Proteomics* **2017**, *17* (20). <https://doi.org/10.1002/pmic.201700011>.
- (8) Ni, M.-W.; Wang, L.; Chen, W.; Mou, H.-Z.; Zhou, J.; Zheng, Z.-G. Modified Filter-Aided Sample Preparation (FASP) Method Increases Peptide and Protein Identifications for Shotgun Proteomics. *Rapid Commun Mass Spectrom* **2017**, *31* (2), 171-178. <https://doi.org/10.1002/rcm.7779>.
- (9) HaileMariam, M.; Eguez, R. V.; Singh, H.; Bekele, S.; Ameni, G.; Pieper, R.; Yu, Y. S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *J Proteome Res* **2018**, *17* (9), 2917-2924. <https://doi.org/10.1021/acs.jproteome.8b00505>.
- (10) Ludwig, K. R.; Schroll, M. M.; Hummon, A. B. Comparison of In-Solution, FASP, and S-Trap Based Digestion Methods for Bottom-Up Proteomic Studies. *J Proteome Res* **2018**, *17* (7), 2480-2490. <https://doi.org/10.1021/acs.jproteome.8b00235>.
- (11) Doellinger, J.; Schneider, A.; Hoeller, M.; Lasch, P. Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-Free Protocol for Proteomics Based on Acid Extraction. *Mol Cell Proteomics* **2020**, *19* (1), 209-222. <https://doi.org/10.1074/mcp.TIR119.001616>.

- (12) Meier, F.; Brunner, A.-D.; Koch, S.; Koch, H.; Lubeck, M.; Krause, M.; Goedecke, N.; Decker, J.; Kosinski, T.; Park, M. A.; Bache, N.; Hoerning, O.; Cox, J.; Räther, O.; Mann, M. Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* **2018**, *17* (12), 2534–2545. <https://doi.org/10.1074/mcp.TIR118.000900>.
- (13) Krieger, J. R.; Wybenga-Groot, L. E.; Tong, J.; Bache, N.; Tsao, M. S.; Moran, M. F. Evosep One Enables Robust Deep Proteome Coverage Using Tandem Mass Tags While Significantly Reducing Instrument Time. *J Proteome Res* **2019**, *18* (5), 2346–2353. <https://doi.org/10.1021/acs.jproteome.9b00082>.
- (14) Bekker-Jensen, D. B.; Martínez-Val, A.; Steigerwald, S.; Rütger, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol Cell Proteomics* **2020**, *19* (4), 716–729. <https://doi.org/10.1074/mcp.TIR119.001906>.
- (15) Diamant, B. J.; Noble, W. S. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. *J Proteome Res* **2011**, *10* (9), 3871–3879. <https://doi.org/10.1021/pr101196n>.
- (16) David, M.; Fertin, G.; Rogniaux, H.; Tessier, D. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *J Proteome Res* **2017**, *16* (8), 3030–3038. <https://doi.org/10.1021/acs.jproteome.7b00308>.
- (17) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nature Methods* **2017**, *14* (5), 513–520. <https://doi.org/10.1038/nmeth.4256>.
- (18) Bittremieux, W.; Meysman, P.; Noble, W. S.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J Proteome Res* **2018**, *17* (10), 3463–3474. <https://doi.org/10.1021/acs.jproteome.8b00359>.
- (19) Li, C.; Li, K.; Chen, T.; Zhu, Y.; He, Q. SW-Tandem: A Highly Efficient Tool for Large-Scale Peptide Identification with Parallel Spectrum Dot

- Product on Sunway TaihuLight. *Bioinformatics* **2019**, 35 (19), 3861–3863. <https://doi.org/10.1093/bioinformatics/btz147>.
- (20) Yu, F.; Haynes, S. E.; Teo, G. C.; Avtonomov, D. M.; Polasky, D. A.; Nesvizhskii, A. I. Fast Quantitative Analysis of TimsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell Proteomics* **2020**, 19 (9), 1575–1585. <https://doi.org/10.1074/mcp.TIR120.002048>.
- (21) Pasa-Tolić, L.; Masselon, C.; Barry, R. C.; Shen, Y.; Smith, R. D. Proteomic Analyses Using an Accurate Mass and Time Tag Strategy. *Biotechniques* **2004**, 37 (4), 621–624, 626–633, 636 passim. <https://doi.org/10.2144/04374RV01>.
- (22) Liu, T.; Belov, M. E.; Jaitly, N.; Qian, W.-J.; Smith, R. D. Accurate Mass Measurements in Proteomics. *Chem. Rev.* **2007**, 107 (8), 3621–3653. <https://doi.org/10.1021/cr068288j>.
- (23) Masselon, C. D.; Kieffer-Jaquinod, S.; Brugière, S.; Dupierris, V.; Garin, J. Influence of Mass Resolution on Species Matching in Accurate Mass and Retention Time (AMT) Tag Proteomics Experiments. *Rapid Commun Mass Spectrom* **2008**, 22 (7), 986–992. <https://doi.org/10.1002/rcm.3447>.
- (24) Zimmer, J. S. D.; Monroe, M. E.; Qian, W.-J.; Smith, R. D. Advances in Proteomics Data Analysis and Display Using an Accurate Mass and Time Tag Approach. *Mass Spectrom Rev* **2006**, 25 (3), 450–482. <https://doi.org/10.1002/mas.20071>.
- (25) Mao, Y.; Zamdborg, L.; Kelleher, N. L.; Hendrickson, C. L.; Marshall, A. G. Identification of Phosphorylated Human Peptides by Accurate Mass Measurement Alone. *Int J Mass Spectrom* **2011**, 308 (2–3), 357–361. <https://doi.org/10.1016/j.ijms.2011.08.006>.
- (26) Ivanov, M. V.; Tarasova, I. A.; Levitsky, L. I.; Solovyeva, E. M.; Pridatchenko, M. L.; Lobas, A. A.; Bubis, J. A.; Gorshkov, M. V. MS/MS-Free Protein Identification in Complex Mixtures Using Multiple Enzymes with Complementary Specificity. *J. Proteome Res.* **2017**, 16 (11), 3989–3999. <https://doi.org/10.1021/acs.jproteome.7b00365>.
- (27) Ivanov, M. V.; Bubis, J. A.; Gorshkov, V.; Tarasova, I. A.; Levitsky, L. I.; Lobas, A. A.; Solovyeva, E. M.; Pridatchenko, M. L.; Kjeldsen, F.; Gorshkov, M. V. DirectMS1: MS/MS-Free Identification of 1000 Proteins

- of Cellular Proteomes in 5 Minutes. *Anal Chem* **2020**, 92 (6), 4326–4333. <https://doi.org/10.1021/acs.analchem.9b05095>.
- (28) Krokhin, O. V.; Spicer, V. Predicting Peptide Retention Times for Proteomics. *Curr Protoc Bioinformatics* **2010**, Chapter 13, Unit 13.14. <https://doi.org/10.1002/0471250953.bi1314s31>.
- (29) Tarasova, I. A.; Masselon, C. D.; Gorshkov, A. V.; Gorshkov, M. V. Predictive Chromatography of Peptides and Proteins as a Complementary Tool for Proteomics. *Analyst* **2016**, 141 (16), 4816–4832. <https://doi.org/10.1039/c6an00919k>.
- (30) Moruz, L.; Tomazela, D.; Käll, L. Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics. *J. Proteome Res.* **2010**, 9 (10), 5209–5216. <https://doi.org/10.1021/pr1005058>.
- (31) Ma, C.; Ren, Y.; Yang, J.; Ren, Z.; Yang, H.; Liu, S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal Chem* **2018**, 90 (18), 10881–10888. <https://doi.org/10.1021/acs.analchem.8b02386>.
- (32) 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>.
- (33) DeepLC can predict retention times for peptides that carry as-yet unseen modifications | bioRxiv <https://www.biorxiv.org/content/10.1101/2020.03.28.013003v1> (accessed Oct 28, 2020).
- (34) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 3146–3154.
- (35) Buryakov, I. A.; Krylov, E. V.; Nazarov, E. G.; Rasulev, U. Kh. A New Method of Separation of Multi-Atomic Ions by Mobility at Atmospheric Pressure Using a High-Frequency Amplitude-Asymmetric Strong

- Electric Field. *International Journal of Mass Spectrometry and Ion Processes* **1993**, 128 (3), 143–148. [https://doi.org/10.1016/0168-1176\(93\)87062-W](https://doi.org/10.1016/0168-1176(93)87062-W).
- (36) Swearingen, K. E.; Moritz, R. L. High-Field Asymmetric Waveform Ion Mobility Spectrometry for Mass Spectrometry-Based Proteomics. *Expert Rev Proteomics* **2012**, 9 (5), 505–517. <https://doi.org/10.1586/epr.12.50>.
- (37) Bonneil, E.; Pfammatter, S.; Thibault, P. Enhancement of Mass Spectrometry Performance for Proteomic Analyses Using High-Field Asymmetric Waveform Ion Mobility Spectrometry (FAIMS). *J Mass Spectrom* **2015**, 50 (11), 1181–1195. <https://doi.org/10.1002/jms.3646>.
- (38) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res.* **2019**, 47 (D1), D442–D450. <https://doi.org/10.1093/nar/gky1106>.
- (39) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, 24 (21), 2534–2536. <https://doi.org/10.1093/bioinformatics/btn323>.
- (40) Teleman, J.; Chawade, A.; Sandin, M.; Levander, F.; Malmström, J. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. *J. Proteome Res.* **2016**, 15 (7), 2143–2151. <https://doi.org/10.1021/acs.jproteome.6b00016>.
- (41) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, 4 (3), 207–214. <https://doi.org/10.1038/nmeth1019>.
- (42) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large

- Proteomic Data Sets. *Mol. Cell Proteomics* **2015**, 14 (9), 2394–2404.
<https://doi.org/10.1074/mcp.M114.046995>.
- (43) Levitsky, L. I.; Ivanov, M. V.; Lobas, A. A.; Gorshkov, M. V. Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach. *J. Proteome Res.* **2017**, 16 (2), 393–397.
<https://doi.org/10.1021/acs.jproteome.6b00144>.
- (44) Meek, J. L. Prediction of Peptide Retention Times in High-Pressure Liquid Chromatography on the Basis of Amino Acid Composition. *Proc Natl Acad Sci U S A* **1980**, 77 (3), 1632–1636.
<https://doi.org/10.1073/pnas.77.3.1632>.
- (45) Smith, R.; Ventura, D.; Prince, J. T. LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review. *Brief Bioinform* **2015**, 16 (1), 104–117. <https://doi.org/10.1093/bib/bbt080>.