

## ACCELERATED ARTICLE

# Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries

Bart Van Puyvelde, Sander Willems, Ralf Gabriels, Simon Daled, Laura De Clerck, Sofie Vande Castele, An Staes, Francis Impens, Dieter Deforce, Lennart Martens, Sven Degroeve, and Maarten Dhaenens\*

**Data-independent acquisition (DIA) generates comprehensive yet complex mass spectrometric data, which imposes the use of data-dependent acquisition (DDA) libraries for deep peptide-centric detection. Here, it is shown that DIA can be redeemed from this dependency by combining predicted fragment intensities and retention times with narrow window DIA. This eliminates variation in library building and omits stochastic sampling, finally making the DIA workflow fully deterministic. Especially for clinical proteomics, this has the potential to facilitate inter-laboratory comparison.**

With data-independent acquisition (DIA), an MS instrument regularly measures precursor ions and continuously cycles through predefined  $m/z$  ratio windows to equally regularly measure the intensity of their fragment ions throughout an LC gradient. This is both more qualitative and quantitative than data-dependent acquisition (DDA), where precursor ions are measured intermittently while fragment ions are only measured stochastically. However, the complexity of DIA data has shown to be very challenging.

To date, the most common peptide-centric way to address this complexity is using previously identified peptides from DDA as targets in the DIA data.<sup>[1]</sup> First, DDA peptide identifications are translated into a spectral library with Peptide Query Parameters (PQPs), which typically contain the sequence as well as the analytical coordinates ( $m/z$ , intensity, and retention time or RT) for the observed ions for a given peptide. These PQPs are then used to compute an evidence score for each target peptide, based on its fragment traces in DIA.<sup>[2]</sup> Ultimately, these evidence scores are supplemented with additional features, e.g., ppm and RT errors, allowing a semi-supervised machine learning algorithm to weigh and re-score the target peptides to obtain a maximum of true targets at an empirically determined false discovery rate (FDR) using the target–decoy approach.<sup>[3–5]</sup>

Unfortunately, deriving PQPs from DDA data intrinsically means transferring its limitations. In fact, fractionation, stochastic data acquisition, processing, and identification introduce bias in the library and require considerable effort. This compromises inter-laboratory comparison and can even alter the biological conclusions between laboratories.<sup>[6]</sup> However, thanks to the availability of state-of-the-art prediction algorithms, these PQPs can now be predicted directly, setting the stage for much easier and much more reproducible peptide-centric DIA data extraction.<sup>[7–9]</sup>

Here, we compare the effect of using libraries from different origins on peptide-centric approaches, by assessing their qualitative and quantitative performance on a public wide window (10–20  $m/z$ ) DIA dataset of HeLa cells<sup>[10]</sup> (**Figure 1**). Three basic spectral libraries were used here, with PQPs derived from a) an experimental DDA dataset, b) a protein sequence database (FASTA), and c) a predicted spectral dataset. Each of these three libraries can be used directly as a source library, or can be converted into a DIA library by using them first on a narrow window (2  $m/z$ ) DIA dataset of the sample. The resulting six possible libraries can all be used alike by the EncyclopeDIA software to identify and quantify wide window DIA data.<sup>[10]</sup> We define in the further text (i) peptide detections as being reported by the software above 1% FDR, (ii) peptidofragments as having deconvoluting charge states and (iii) robust peptides as being detected in three separate runs with at least three transitions.

In-house or public DDA source libraries are frequently built by extensive fractionation of samples. With adequate statistical

B. Van Puyvelde, S. Willems, S. Daled, Dr. L. De Clerck, S. Vande Castele, Prof. D. Deforce, Dr. M. Dhaenens

ProGenTomics

Laboratory of Pharmaceutical Biotechnology

Ghent University

9000 Ghent, Belgium

E-mail: maarten.dhaenens@ugent.be

R. Gabriels, A. Staes, Prof. F. Impens, Prof. L. Martens, Prof. S. Degroeve

VIB-UGent Center for Medical Biotechnology

9000 Ghent, Belgium

R. Gabriels, A. Staes, Prof. F. Impens, Prof. L. Martens, Prof. S. Degroeve

Department of Biomolecular Medicine


Ghent University

9000 Ghent, Belgium

A. Staes, Prof. F. Impens

VIB Proteomics Core

9000 Ghent, Belgium

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/pmic.201900306>

© 2020 The Authors. *Proteomics* published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/pmic.201900306

control, such proteotypic libraries allow direct peptide detections in wide window DIA (Figure 1Aa).<sup>[11]</sup> We illustrate this by using the publically available Pan-Human library, which contains nearly 10 000 proteins derived from 331 DDA runs on a range of human cell lines and tissues<sup>[12]</sup> (Figure 1Ba). To reduce the effort and variability from DDA library building, a library-free peptide-centric data analysis workflow was proposed recently.<sup>[13]</sup> Herein, the PECAN (or Walnut) scoring algorithm allows direct detection of peptides derived from a FASTA in wide window DIA data (Figure 1Ab). This is akin to a source library that i) contains only peptide sequences and  $m/z$  coordinates, and ii) lacks prior selection of proteotypic peptides. On wide window DIA data, this approach thus provides a limited number of PQPs, which is not sufficient to differentiate between the high number of false targets, i.e., true negatives, and the lower number of true positives in the library.<sup>[14]</sup> This manifests as indiscernible target and decoy score distributions, resulting in a very high false negative rate (FNR; Figure 1Bb).

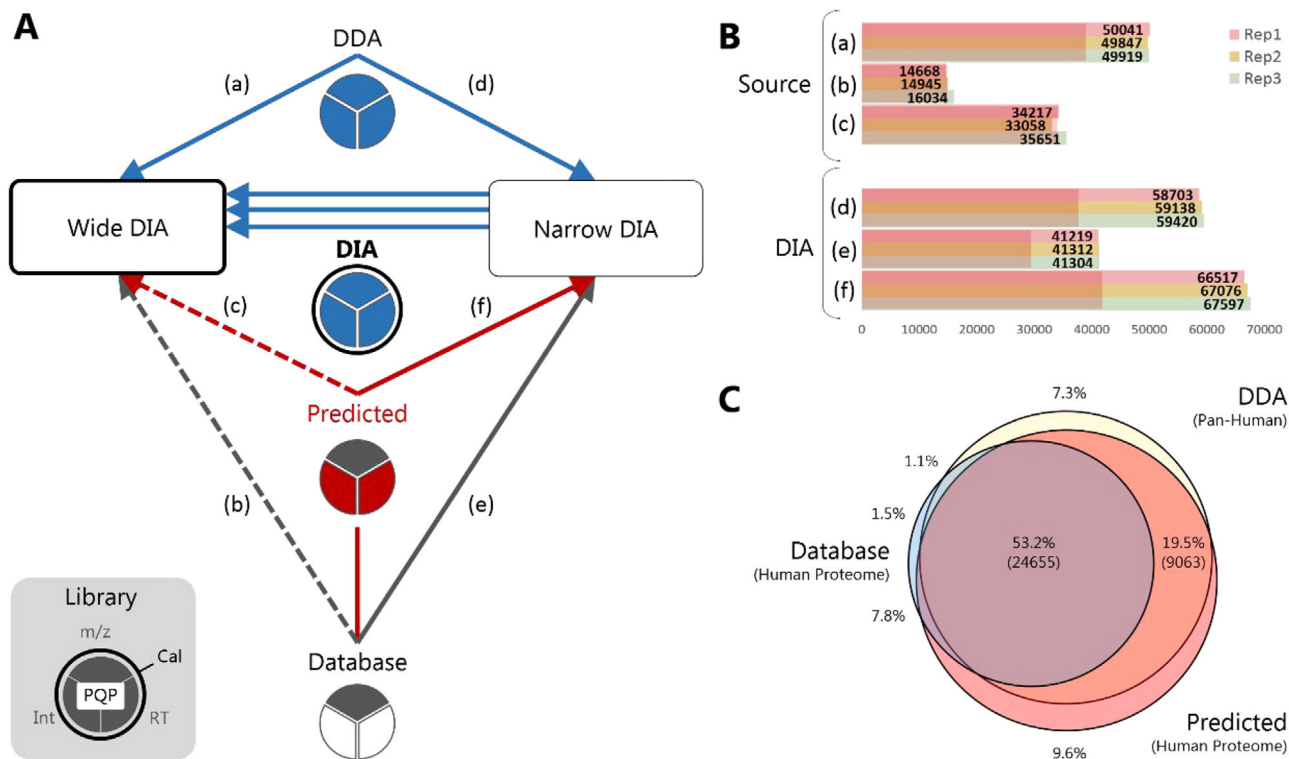
Here, we propose a promising way to improve upon the FASTA source library—while still omitting prior DDA—by predicting fragment ion intensity and RT in silico (Figure 1Ac; Figures S1 and S2, Supporting Information). Using a spectral dataset with such predicted fragment intensities (MS<sup>2</sup>PIP) and peptide RTs (Elude) more than doubles the number of peptides detected in the wide window DIA (Figure 1Bc).<sup>[7,15]</sup> However, considering all tryptic peptides in a Human proteome still underperforms compared to the Pan-Human DDA library, which is fully contained in the predicted spectral dataset (Figure 1Ba,Bc). Notably, this is not due to poor prediction because predicting only those peptides present in the Pan-Human library performs very similar to using the Pan-Human library directly (Figure S3, Supporting Information) and the underperformance can thus only be attributed to the many false targets when using the complete database.<sup>[11]</sup> An elegant way to filter out false target peptides upfront, is by measuring a pool from every condition with staggered narrow window DIA (Figure 1Ad–f). This reduces MS<sup>2</sup> chimericity to DDA-like quality in a DIA setting, allowing detection with increased specificity. This accurate prior filtering makes the statistical burden of false targets in the wide window DIA surmountable again. Notably, due to instrument limitations this “Precursor Acquisition Independent From Ion Count” (PACIFIC)<sup>[16]</sup> can currently only be performed by means of gas phase fractionation (GPF), i.e., sampling different  $m/z$  regions separately.<sup>[10]</sup> Still, the added acquisition depth and specificity allows for 88k (DDA), 47k (FASTA), and 95k (predicted) doubly and triply charged peptide detections as reported by the software, corresponding to 84k, 44k, and 90k peptidofoms in six narrow window GPF DIA runs of a HeLA cell lysate (Figure S4, Supporting Information). To assure that this additional filtering is accurate, we confirmed the estimated FDR by using an entrapment experiment wherein we included *Pyrococcus furiosus* proteins as false targets alongside the expected human proteins in the respective source libraries.<sup>[17]</sup> Hereby, the measured FDR for narrow window DIA filtering is 2% for the DDA, 1% for the FASTA, and 1% for the predicted source library, in accordance with the theoretically estimated FDR based on the target-decoy strategy. In the process, we can measure the identification cost of adding false targets: adding 3–6% false targets results in an

## Significance Statement

Data-independent acquisition (DIA) is quickly developing into the most comprehensive strategy to analyse a sample on a mass spectrometer. Correspondingly, a wave of data analysis strategies has followed suit, improving the yield from DIA experiments with each iteration. As a result, a worldwide wave of investments in DIA is already taking place in anticipation of clinical applications. Yet, there is considerable confusion about the most useful and efficient way to handle DIA data, given the plethora of possible approaches with little regard for compatibility and complementarity. In our study, we outline the currently available peptide-centric DIA data analysis strategies in a unified graphic called the DIAMOND DIAGRAM. This leads us to an innovative and easily adoptable approach based on predicted spectral information. Most importantly, our contribution removes what is arguably the biggest bottleneck in the field: the current need for data-dependent acquisition (DDA) prior to DIA analysis. Fractionation, stochastic data acquisition, processing, and identification all introduce bias in the library. By generating libraries through data independent, i.e., deterministic acquisition, stochastic sampling in the DIA workflow is now fully omitted. This is a crucial step toward increased standardization. Additionally, our results demonstrate that a proteome-wide predicted spectral library can surrogate an exhaustive DDA Pan-Human library that was built based on 331 prior DDA runs.

average decrease of 1–2% in detections (see Entrapment Section in Supporting Information Methods).

Additionally, the peptide detections in narrow window DIA can be translated into novel and integrated PQPs, which are calibrated to the specific LC–MS system and are specific to DIA (Figure 1A). This approach was recently made readily applicable as chromatogram libraries: DIA libraries of narrow window DIA peptide detections comprising their calibrated PQPs.<sup>[10]</sup> Such chromatogram libraries outperform direct wide window DIA extraction for every source library. The modest gain for a DDA source library ( $\approx 20\%$ ) derives mainly from PQP calibration, as only 50% of the source peptides was filtered out (Figure 1Ba,Bd). In contrast, in the FASTA source library, 98.5% of the peptides were filtered out, and RT and intensity coordinates were generated de novo. Taken together, this resulted in the largest gain ( $\approx 170\%$ ; Figure 1Bb,Be). Finally, the chromatogram library derived from a predicted spectral library increases the number of detections by  $\approx 100\%$  compared to direct wide window DIA data extraction, making it the most efficient overall peptide detection strategy of the DIAMOND DIAGRAM (Figure 1Bc,Bf). Importantly, when looking only at robust peptide detections, i.e., with a minimum of three transitions and found in triplicate, the gain compared to the Pan-Human library is rather modest. Additionally, the robust peptides detected by all three chromatogram libraries show a large overlap, convincingly showing that the Pan-Human library is very exhaustive and that all three chromatogram libraries mainly detect proteotypic peptides



**Figure 1.** Peptide-centric data extraction from wide window DIA data. **A**) DIAMOND DIAgram presenting peptide-centric strategies for DIA data extraction. Peptide-centric approaches rely on libraries (central column) that contain Peptide Query Parameters (PQPs), which are derived from the peptide sequence and can additionally contain the three ion coordinates, i.e., mass to charge ratio ( $m/z$ ), Intensity (Int), and retention time (RT) (three-part pie charts). These can either be experimental (blue), theoretical (grey), or predicted (red). PQPs are used to score the evidence of peptide detections in continuous DIA data (boxes). These are supplemented with additional features of the match so that a support vector machine can weigh and re-score them to obtain a maximum of true targets at an empirically determined FDR using the target-decoy approach (arrow heads). DDA source libraries (both in-house and public) only comprise prior proteotypic peptide identifications and contain measured PQPs for all three ion coordinates. These are therefore directly applicable to quantify peptides in 10–20  $m/z$  wide window DIA (Wide DIA) data (a). However, when a proteome FASTA is used as a source library, sensitivity is reduced (dashed arrow), i.e., too many false negatives are produced due to the high statistical burden (b). This also holds for libraries with predicted fragment intensities (MS<sup>2</sup>PIP) and RT (Elude), albeit to a lesser extent (c). Prior 2  $m/z$  narrow window DIA (Narrow DIA) provides the specificity to remove false targets in the sample first (d–f). The DIA ion coordinates from these detections can additionally be integrated into new and calibrated PQPs (cal). These DIA libraries, called chromatogram libraries, can be derived from any source library (triple arrow). **B**) Doubly and triply charged peptide detections in wide window DIA following each of the routes depicted in (A). Shading highlights the number of robust peptides that is detected in triplicate wide window DIA runs with at least three transitions, allowing robust quantification. **C**) Comparison of the identified robust peptides in Wide DIA for route (d–f). The large overlap shows that all three approaches detect proteotypic peptides. Only peptides of double and triple charge that are detected in triplicate wide window DIA runs with at least three transitions are shown.

(Figure 1C). Peptides unique to the Pan-Human library include very high molecular masses that were not predicted, high molecular weight peptides that generate many doubly charged transitions that are not predicted by default, as well as very small peptides with inherently poor RT or fragmentation pattern predictions. Peptides that are unique to the predicted library are all peptides that were not present in the Pan-Human source library and are very low abundant in the wide window DIA data, implying they were missed during the DDA sampling in the Pan Human library (Figure S4, Supporting Information). Note that some peptides will pass the detection threshold only in the narrow window DIA and not in the wide window DIA because of increased interference in the latter (1788 for the predicted and 673 for the Pan-Human). Importantly, the PQP requirements of the source library for building chromatogram libraries on narrow window DIA are relatively liberal: the measured Pan-Human library was acquired on a TripleTOF instrument but allows wide

window DIA data peptide detection on an Orbitrap instrument. The in silico equivalent is that 95% of the detected peptides overlap when the MS<sup>2</sup>PIP engine is trained on either Orbitrap or TripleTOF data. As a result, other fragment ion intensity predictors such as ProSIT and Deep Mass<sup>[8,9]</sup> perform similarly when combined with narrow window DIA<sup>[18]</sup> (Figures S5 and S6, Supporting Information). Overall, the peptide-centric workflow seems to have matured to a level that has covered much of the most obvious growing potential. Fortunately, very different ways of mining DIA data are continuously being presented, such as the use of neural networks or building ion networks.<sup>[19,20]</sup>

We conclude that predicted libraries are highly relevant and performant for wide window DIA identification, and that three elements of a spectral library affect its overall performance: i) the amount of false targets included, ii) the amount of informative PQPs, and iii) the accuracy of PQPs on the specific instrument setup. In this study, we could show that a narrow window DIA

Received: September 19, 2019

Revised: December 20, 2019

Published online:

acquisition of six GPFs combined with a predicted spectral library of the full human proteome was able to surrogate a measured DDA Pan-Human library, thus liberating the DIA workflow from any stochastic acquisition. Especially for clinical proteomics, this can facilitate inter-laboratory comparison. Importantly, the software tools MS<sup>2</sup>PIP, ELUDE, and EncyclopeDIA are all instrument independent, publicly available, and mutually compatible, thus making this workflow immediately accessible to everybody interested.

## Code Availability

MS<sup>2</sup>PIP, Elude, Prosit, and EncyclopeDIA are open source, licensed under the Apache-2.0 License, and are hosted on [https://github.com/compomics/ms2pip\\_c](https://github.com/compomics/ms2pip_c), <https://github.com/percolator/percolator>, <https://github.com/kusterlab/prosit>, and <https://bitbucket.org/searleb/encyclopedia/wiki/Home>. All supporting material is available on <https://github.com/brvpuyve/MS2PIP-for-DIA/>.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

This research was mainly funded by mandates from the Research Foundation Flanders (FWO) awarded to B.V.P. (grant number 11B4518N), R.G. (grant number 1S50918N), and M.D. (12E9716N). Partial funding was received through project grants from the FWO (G013916N and G042518N), from the European Union's Horizon 2020 Program under Grant Agreement 823839 (H2020-INFRAIA-2018-1), and from a Ph.D. grant from the Flanders Agency Entrepreneurship and Innovation (VLAIO) awarded to LDC (SB-141209).

## Author Contributions

B.V.P., S.W., and R.G. contributed equally to this work. B.V.P. performed all data analysis at the ProGenTomics facilities. The initial experimental design was conceived and performed by B.V.P., S.W., M.D., S.Da., L.D.C., A.S., D.D., and F.I. R.G., S.De., and L.M. performed all machine learning predictions. M.D., B.V.P., R.G., and S.W. wrote the draft manuscript. All authors provided critical feedback during research and writing. M.D. conceived the idea of using predicted libraries for DIA data extraction and supervised the project.

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

bioinformatics, data-independent acquisition, label-free quantification, peptide-centric

- [1] Y. S. Ting, J. D. Egertson, S. H. Payne, S. Kim, B. MacLean, L. Kall, R. Aebersold, R. D. Smith, W. S. Noble, M. J. MacCoss, *Mol Cell Proteomics* **2015**, *14*(9), 2301.
- [2] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B. C. Collins, R. Aebersold, *Mol. Syst. Biol.* **2018**, *8*, 8126.
- [3] L. Reiter, O. Rinner, P. Picotti, R. Huttenhain, M. Beck, M. Y. Brusniak, M. O. Hengartner, R. Aebersold, *Nat. Methods* **2011**, *8*, 430.
- [4] L. Kall, J. D. Canterbury, J. Weston, W. S. Noble, M. J. MacCoss, *Nat. Methods* **2007**, *4*, 923.
- [5] J. Teleman, H. L. Röst, G. Rosenberger, U. Schmitt, L. Malmström, J. Malmström, F. Levander, *Bioinformatics* **2015**, *31*, 555.
- [6] E. Govaert, K. Van Steendam, S. Willems, L. Vossaert, M. Dhaenens, D. Deforce, *Proteomics* **2017**, *17*, 1700052.
- [7] R. Gabriels, L. Martens, S. Degroeve, *Nucleic Acids Res.* **2019**, *47*, W295.
- [8] S. Gessulat, T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, U. Reimer, H.-C. Ehrlich, S. Aiche, B. Kuster, M. Wilhelm, *Nat. Methods* **2019**, *16*, 509.
- [9] S. Tiwary, R. Levy, P. Gutenbrunner, F. Salinas Soto, K. K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic, J. Cox, *Nat. Methods* **2019**, *16*, 519.
- [10] B. C. Searle, L. K. Pino, J. D. Egertson, Y. S. Ting, R. T. Lawrence, B. X. MacLean, J. Villén, M. J. MacCoss, *Nat. Commun.* **2018**, *9*, 5128.
- [11] G. Rosenberger, I. Bludau, U. Schmitt, M. Heusel, C. L. Hunter, Y. Liu, M. J. MacCoss, B. X. MacLean, A. I. Nesvizhskii, P. G. A. Pedrioli, L. Reiter, H. L. Rost, S. Tate, Y. S. Ting, B. C. Collins, R. Aebersold, *Nat. Methods* **2017**, *14*, 921.
- [12] G. Rosenberger, C. C. Koh, T. Guo, H. L. Röst, P. Kouvonen, B. C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, M. Faini, O. T. Schubert, P. Faridi, H. A. Ebhardt, M. Matondo, H. Lam, S. L. Bader, D. S. Campbell, E. W. Deutsch, R. L. Moritz, S. Tate, R. Aebersold, *Sci. Data* **2014**, *1*, 140031.
- [13] Y. S. Ting, J. D. Egertson, J. G. Bollinger, B. C. Searle, S. H. Payne, W. S. Noble, M. J. MacCoss, *Nat. Methods* **2017**, *14*, 903.
- [14] N. Colaert, S. Degroeve, K. Helsens, L. Martens, *J. Proteome Res.* **2011**, *10*, 5555.
- [15] L. Moruz, A. Staes, J. M. Foster, M. Hatzou, E. Timmerman, L. Martens, L. Kall, *Proteomics* **2012**, *12*, 1151.
- [16] A. Panchoaud, A. Scherl, S. A. Shaffer, P. D. Von Haller, H. D. Kulasekara, S. I. Miller, D. R. Goodlett, *Anal. Chem.* **2009**, *81*, 6481.
- [17] M. Vaudel, J. M. Burkhardt, D. Breiter, R. P. Zahedi, A. Sickmann, L. Martens, *J. Proteome Res.* **2012**, *11*, 5065.
- [18] B. C. Searle, K. E. Swearingen, C. A. Barnes, T. Schmidt, S. Gessulat, B. Kuster, M. Wilhelm, *bioRxiv* **2019**, 682245.
- [19] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, M. Ralser, *Nature Methods* **2020**, *17*, 41.
- [20] S. Willems, S. Daled, B. Van Puyvelde, L. De Clerck, S. Vande Castele, F. Van Nieuwerburgh, D. Deforce, M. Dhaenens, *bioRxiv* **2019**, <https://doi.org/10.1101/726273>.

# PROTEOMICS

**Supporting Information**

**for Proteomics**

**DOI 10.1002/pmic.201900306**

Bart Van Puyvelde, Sander Willems, Ralf Gabriels, Simon Daled, Laura De Clerck, Sofie Vande Castele, An Staes, Francis Impens, Dieter Deforce, Lennart Martens, Sven Degroeve and Maarten Dhaenens

**Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries**