The future of peptide-centric Data-Independent Acquisition is predicted

Van Puyvelde, B.*[a], Willems, S.*[a], Gabriels, R.*[b,c], Daled, S.[a], De Clerck, L.[a], Staes, A.[b,c,d], Impens, F.[b,c,d], Deforce, D.[a], Martens, L.[b,c], Degroeve, S.[b,c], Dhaenens, M.[§a]

*: Authors contributed equally

a: ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, Belgium

b: VIB-UGent Center for Medical Biotechnology, Ghent, Belgium

c: Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

d: VIB Proteomics Core, Ghent, Belgium

§: Corresponding author; Telephone: +32 (0)9 264 83 56; E-mail: maarten.dhaenens@ugent.be

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. We here show that DIA can be redeemed from this dependency by combining predicted fragment intensities and retention times with narrow window DIA.

With DIA, an MS instrument regularly measures precursor ions and continuously cycles through predefined mass over charge ratio (*m/z*) windows to equally regularly measure the intensity of their fragment ions throughout a liquid chromatography (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 way to address this complexity is using previously identified peptides from DDA as targets in the DIA data. 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 (1). Ultimately, these evidence scores are supplemented with additional features, e.g. ppm and RT errors, allowing a support vector machine (SVM) 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 (2)(3).

Unfortunately, deriving PQPs from DDA data intrinsically means transferring its limitations. In fact, raw data acquisition, processing and identification introduce bias and require considerable effort. 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 (4)(5)(6).

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 (7) (**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 (7).

In-house or public DDA source libraries are frequently built by extensive fractionation of samples. With adequate statistical control, such proteotypic libraries allow direct peptide detections in wide window DIA (**Figure 1Aa**) (8). 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 (9) (**Figure 1Ba**). To reduce the effort and variability from DDA library building, a library-free peptide-centric data analysis workflow was proposed recently (10). 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 (11). This manifests as indiscernible target and decoy score distributions, resulting in a very high False Negative Rate (FNR) (**Figure 1Bb**).

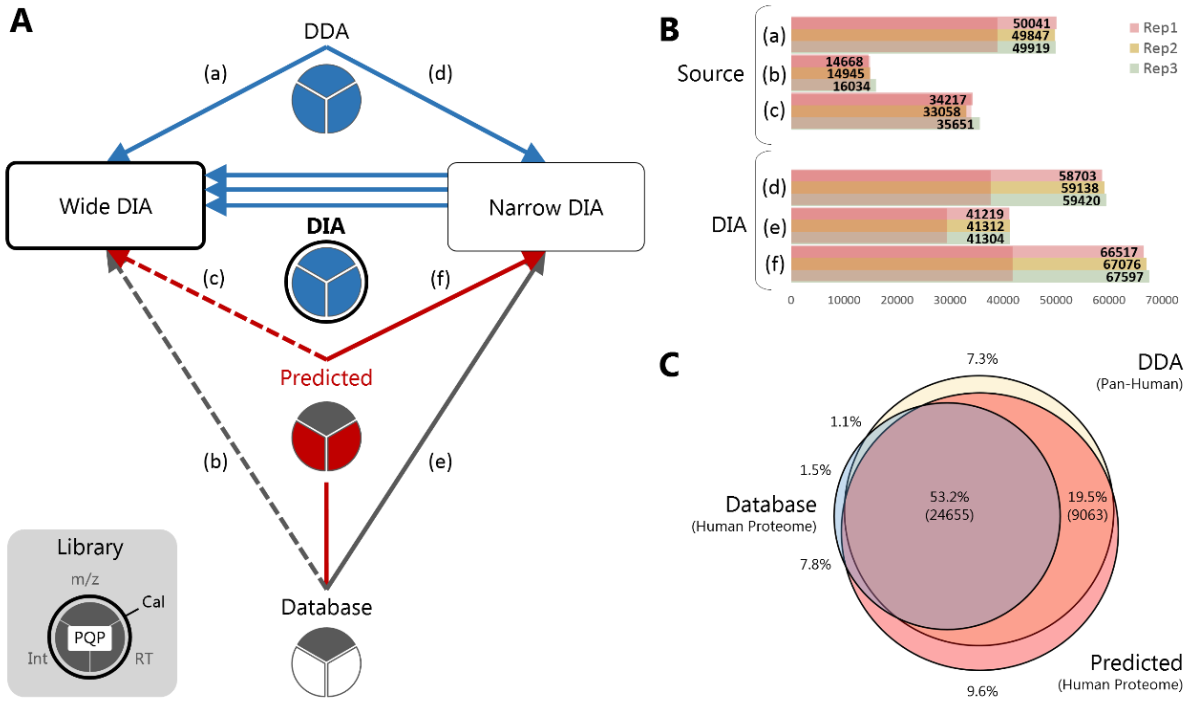

Figure 1. Peptide-centric data extraction from wide window DIA data. (A) DIAmond DIAgram presenting peptide-centric strategies for DIA data extraction. Peptide-centric approaches rely on libraries **(central column)** that contain Peptide Query Parameters (PQPs) which are derived from the peptide sequence 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 (MS²PIP) 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)(e)(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)** Peptide detections in wide window DIA following each of the routes depicted in **(A).** Shading highlights the number of peptides that is detected in triplicate wide window DIA runs with at least three transitions, allowing robust quantification. **(C)** Comparison of the identified peptide sequences in Wide DIA for route **(d), (e)** and **(f)**. The large overlap shows that all three approaches detect proteotypic peptides. Only peptides that are detected in triplicate wide window DIA runs with at least three transitions are shown.

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**). Using a spectral dataset with such predicted fragment intensities (MS²PIP) and peptide RTs (Elude) more than doubles the number of peptides detected in the wide window DIA (**Figure 1Bc**) (4)(12). 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 and 1Bc**). Notably, this is not due to poor prediction (see MS²PIP section in Methods) and can thus only be attributed to the many false targets when using the complete database (8). 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, 1Ae and 1Af**). This reduces MS2 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) (13) can currently only be performed by means of gas phase fractionation (GPF), i.e. sampling different *m/z* regions separately (7). Still, the added acquisition depth and specificity allows the detection of 91k (DDA), 47k (FASTA) and 95k (predicted) peptides in six narrow window GPF DIA runs of a HeLA cell lysate. 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 (14). 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 average decrease of 1-2% in detections (see Entrapment section in Methods).

Additionally, the peptide detections in narrow window DIA can be translated into novel and integrated PQPs, which are calibrated to the specific LCMS system and 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 (7). Such chromatogram libraries outperform direct wide window DIA extraction for every source library. The modest gain for a DDA source library (~20%) derives mainly from PQP calibration, as only 50% of the source peptides was filtered out (**Figure 1Ba and 1Bd**). 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 (~170%) (**Figure 1Bb and 1Be**). Finally, the chromatogram library derived from a predicted spectral library increases the number of detections by ~100% compared to direct wide window DIA data extraction, making it the most efficient overall peptide detection strategy of the DIAmond DIAgram (**Figure 1Bc and 1Bf**). The large overlap between the peptide sequences from all three chromatogram libraries convincingly shows that the Pan-Human library is very exhaustive and that all three chromatogram libraries mainly detect proteotypic peptides (**Figure 1C**). 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 MS²PIP engine is trained on either HCD or TripleTOF data. As a result, other fragment ion intensity predictors such as Prosit and Deep Mass (5)(6) are likely to perform similarly.

We therefore 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 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 that was built from 331 DDA runs. Importantly, the software tools MS²PIP, ELUDE

and EncyclopeDIA are all instrument independent, publicly available, and mutually compatible, thus making this workflow immediately accessible to everybody interested.

Code availability

MS²PIP, Elude 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/percolator/percolator and https://bitbucket.org/searleb/encyclopedia/wiki/Home. All supporting material is available on https://github.com/brvpuyve/MS2PIP-for-DIA/.

Acknowledgments

This research was mainly funded by mandates from the Research Foundation Flanders (FWO) awarded to BVP [grant number 11B4518N], RG [grant number 1S50918N] and MD [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 PhD grant from the Flanders Agency Entrepreneurship and Innovation (VLAIO) awarded to LDC [SB-141209]. We thank B. Searle for the valuable discussions and for implementing .msp compatibility into EncyclopeDIA.

Competing interests

The authors have declared no conflict of interest.

References

1. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol. 2018;

2. Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods [Internet]. 2011/03/23. 2011;8(5):430–5. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21423193

3. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods [Internet]. 2007/10/24. 2007;4(11):923–5. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17952086

4. Gabriels R, Martens L, Degroeve S. Updated MS²PIP web server delivers fast and accurate MS² peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. Nucleic Acids Res. 2019;

5. Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods. 2019;

6. Tiwary S, Levy R, Gutenbrunner P, Salinas Soto F, Palaniappan KK, Deming L, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods. 2019;

7. Searle BC, Pino LK, Egertson JD, Ting YS, Lawrence RT, MacLean BX, et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun. 2018;

8. Rosenberger G, Bludau I, Schmitt U, Heusel M, Hunter CL, Liu Y, et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods [Internet]. 2017/08/22. 2017;14(9):921–7. Available from:

https://www.ncbi.nlm.nih.gov/pubmed/28825704

9. Rosenberger G, Koh CC, Guo T, Röst HL, Kouvonen P, Collins BC, et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. Sci Data. 2014;

10. Ting YS, Egertson JD, Bollinger JG, Searle BC, Payne SH, Noble WS, et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods [Internet]. 2017/08/08. 2017;14(9):903–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28783153

11. Colaert N, Degroeve S, Helsens K, Martens L. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res [Internet]. 2011;10(12):5555–61. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21995378

12. Moruz L, Staes A, Foster JM, Hatzou M, Timmerman E, Martens L, et al. Chromatographic retention time prediction for posttranslationally modified peptides. Proteomics [Internet]. 2012/05/12. 2012;12(8):1151–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22577017

13. Panchaud A, Scherl A, Shaffer SA, Von Haller PD, Kulasekara HD, Miller SI, et al. Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. Anal Chem. 2009;

14. Vaudel M, Burkhart JM, Breiter D, Zahedi RP, Sickmann A, Martens L. A complex standard for protein identification, designed by evolution. J Proteome Res [Internet]. 2012;11(10):5065–71. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22489649

METHODS

CONTENTS

Introduction 1
Library Generation 1
DDA 1
Database (FASTA) 2
Predicted 2
DIA 2
RAW file processing 3
DIA data analysis: EncyclopeDIA 3
FDR assessment by entrapment 4
Prediction Models: Elude and MS²PIP 5
Elude: Retention Time prediction 5
MS²PIP: intensity Prediction 6
References 7

Introduction

DIA data has been presented as a permanent record of everything. Thus, applying our novel approach can significantly broaden the biological perspective on newly acquired as well as existing data. Using predicted spectral libraries to replace measured DDA libraries not only reduces workload and increases reproducibility; it will also facilitate the implementation of DIA into more applied fields such as clinical proteomics. Since the software tools MS²PIP, Elude and EncyclopeDIA are instrument independent, publically available and mutually compatible, the presented workflow is accessible to everybody and directly applicable (1–3). Therefore, we present this methods section in the form of a systematic tutorial. Briefly, both source and DIA libraries can be used in EncyclopeDIA to detect peptides in wide window DIA. However, converting source libraries into a DIA library will significantly improve the number of peptides that can be detected. This requires an additional narrow window DIA of several gas phase fractions (GPF) of a mixture of the samples. When these GPFs are acquired in the same batch as the wide window DIA, the benefit of PQP calibration is maximized.

All external resources are available on GitHub https://github.com/brvpuyve/MS2PIP-for-DIA for reproducibility.

Library Generation

DDA

An EncyclopeDIA .dlib version of the Pan-Human spectral library is publicly available on the EncyclopeDIA BitBucket homepage (4). This version contains 211k unique precursors (159k unique peptide sequences). Alternatively, EncyclopeDIA accepts Skyline .BLIB, Spectronaut .csv, MaxQuant msms.txt, .TraML and .msp files.

Database (FASTA)

Using a FASTA database does not require a separate library. More specifically, Walnut (a GUI re-implementation of the PECAN algorithm) is part of EncyclopeDIA and can directly detect peptides from DIA data using a FASTA database(5).

Here, we used the human SwissProt proteome (UP000005640 downloaded on 12 February 2019, 20426 target sequences) downloaded as FASTA. The proteome was concatenated with the iRT FASTA obtained from the Biognosys webpage (on 12 February 2019) (6).

Predicted

Creating a predicted spectral library requires three steps: (i) creating an MS²PIP input PEPREC (peptide record) file from a FASTA, (ii) feeding that file to MS²PIP for predicting intensities and (iii) adding predicted retention times (RT) from Elude. For ease-of-use, we wrapped these three steps into a pipeline (fasta2speclib), that is included in the MS²PIP GitHub Repository.

MS²PIP is accessible either through the web server (https://iomics.ugent.be/MS²PIP/) or via a local installation (https://github.com/compomics/MS²PIP_c/). A local installation is required to use the fasta2speclib pipeline. Here, MS²PIP (version 20190130) was downloaded and installed from the MS²PIP GitHub repository, as described in the extended install instructions. For RT prediction, we employed Elude version 3.02, which is available from the Percolator GitHub repository (https://github.com/percolator/percolator/ releases).

Briefly, the fasta2speclib pipeline makes use of Biopython to read the FASTA and uses Pyteomics for the *in silico* digestion of the protein sequences (7,8). Next, redundant peptides and peptides not meeting the peptide length and precursor mass restrictions are removed from the peptide list. Following this step, all combinations of the requested charge states and modifications are added. Predicted spectra and RTs are then generated for all peptide-charge-modification combinations using MS²PIP and Elude. Finally, the results are written to a spectral library file (.msp, .mgf or .csv). Depending on the computational resources, a full human proteome can be predicted in just a few hours.

The fasta2speclib pipeline can be called through the command line interface as follows:

```
Python “fasta2speclib.py” [-h] [-o OUTPUT_FILENAME] [-c CONFIG_FILENAME] “fasta_filename”
```

The results presented in this manuscript were generated by predicting a spectrum for every 2+ and 3+ tryptic peptide in the aforementioned FASTA, using the pre-trained MS²PIP Orbitrap-HCD model and the Elude RT. These models are described in more detail under “Prediction models”. Only tryptic peptides with a minimum length of 7 amino acid residues and a maximum precursor mass of 5000 Da were considered. Carbamidomethylation and oxidation were set as respectively fixed and variable modification, and two missed cleavages were allowed. The *in silico* spectral library was exported to an .msp file containing 3.3M precursors (between 400 – 1000 *m/z*). In the current version of MS²PIP (v20190624) the RT from Elude is automatically converted into minutes and written on a separate line in the .msp file. These predictions were performed on a Linux operated machine (Intel Xeon CPU X5670, 24 processors, 40 GB RAM) and took four hours.

DIA

DIA libraries, called chromatogram libraries, are generated by interrogating narrow window DIA data with any of the above source libraries. Details are described under “DIA data analysis: EncyclopeDIA”.

RAW file processing

We used the publicly available dataset of the EncyclopeDIA article (MassIVE MSV000082805) of the HeLa S3 lysates to assay the different routes in the DIAmond DIAgram (**Figure 1A, boxes**). The three wide window DIA replicate runs were acquired with 25 overlapping 24 *m/z* windows and the staggered 4 *m/z* narrow window DIA data comprises six gas phase fractions (GPF) of 100 *m/z* each, together covering a 400 - 1000 *m/z* mass range. Following peak picking, these runs were demultiplexed into 12 *m/z* (wide DIA) and 2 *m/z* (narrow DIA) windows, respectively, and converted into mzML output files by MSConvertGUI with following parameters (9,10):

> **Peak picking:** Vendor specific algorithms (algorithms available for all vendors, except Waters)
> **Demultiplexing:** overlap only with a mass error of 10 ppm

DIA data analysis: EncyclopeDIA

We downloaded EncyclopeDIA from bitbucket (https://bitbucket.org/searleb/EncyclopeDIA/downloads/?tab=downloads) (version 0.8.2, 2019-05-21). EncyclopeDIA is a Java application developed to perform narrow- and wide window DIA data analysis. The application can be run on all three major operating systems (Windows, Mac and Linux), but in this project it was used on a Windows 7 operating system (Lenovo Thinkstation, Intel Xeon E5-2620 24 processors, 128 GB ram). EncyclopeDIA was operated through the graphical user interface but also comes with a command-line interface.

For applying EncyclopeDIA on predicted spectral libraries, the .msp file is first converted into a .dlib file using the conversion tool embedded in EncyclopeDIA. EncyclopeDIA also allows the conversion of other spectrum library formats into .dlib files.

General settings in EncyclopeDIA applied to all searches in this project are as follows:

> **Background:** Human_iRT.fasta
> **Target/Decoy approach:** Normal Target/Decoy
> **Data Acquisition Type:** Non-Overlapping DIA (the narrow window DIA was already deconvoluted by MS Convert, therefore EncyclopeDIA does not need to perform an extra deconvolution.)
> **Enzyme:** Trypsin
> **Fragmentation:** CID/HCD (b- and y- fragments)
> **Precursor/Fragment/Library Mass Tolerance:** 10.0 ppm
> **Percolator Version:** v3-01
> **Number of Quantitative Ions:** 5
> **Minimum number of Quantitative Ions:** 3
> **Number of Cores:** 24 (depending on the number of CPU cores you allow/have available)

To allow direct comparison of all six routes of the DIAmond DIAgram, all libraries were trimmed upfront to retain only peptides in the 400 - 1000 *m/z* mass range. For the Pan-Human DDA library this results in 194k precursors, all charge states still included. Approximately 95% of the identified peptides on the wide window DIA were 2+ and 3+ and the other charge states were manually removed from the result file for comparison. The FASTA search was performed using Walnut, considering 2+ and 3+ precursors only. Finally, a third library was predicted by MS²PIP using the same FASTA. All three source libraries were separately used to detect peptides directly in the triplicate wide window HeLa DIA runs (**Figure 1Aa-c**). When the three source libraries were used to search the narrow window DIA data (**Figure 1Ad, 1Ae, 1Af**), this resulted in three DIA-based chromatogram libraries (.elib) of size 91k (DDA), 47k (FASTA) and the 95k (Predicted) peptides, respectively. All three .elibs were used to search the wide window DIA data with the above parameters.

Figure 1B depicts the number of detected peptides in each replicate as reported by EncyclopeDIA. Additionally, the peptide quantification reports were exported as .txt files and peptide sequences with at least 3 transitions and non-zero intensities in all three wide window DIA samples were selected. These are represented as the shaded portion of the bar chart. All log and result files of the searches were exported for future reference and are available on our GitHub repository.

FDR assessment by entrapment

We validated the theoretical FDR from the target-decoy approach during chromatogram library building by performing an entrapment experiment with *Pyrococcus furiosus*. In short, this is a way to additionally validate the target-decoy FDR estimation (11).Only peptides between 400 - 1000 *m/z* were considered and each source library requires a different *P. furiosus* input:

- A public *P. furiosus* dataset acquired on an LTQ-Orbitrap Velos (Thermo Fisher Scientific, Massachusetts, US) was used to supplement the Pan-Human DDA library (ProteomeXchange with identifier PXD001077)(11). Database searching was performed on the resulting .mgf file with Mascot Daemon (version 2.6.1) using following search parameters: a maximum of one missed cleavage, peptide charges 2+ to 4+, peptide mass tolerance of 10 ppm, fragment ion tolerance of 0.5 Da, carbamidomethylation of Cysteine as fixed modification and oxidation of Methionine as variable modification. The resulting .DAT file was parsed into a .BLIB using the Skyline built-in tool BiblioSpec. The .BLIB file was parsed by EncyclopeDIA into a .dlib file. Finally, the resulting .dlib file (5.5k unique precursors) was combined with the already existing Pan-Human .dlib file of 194k peptides using EncyclopeDIA.
- For the FASTA database, we concatenated our FASTA with all 2052 *P. furiosus* UniProt entries (downloaded on June 13, 2018). Walnut parameters for library-free searching were set as described above, meaning that only 2+ and 3+ peptides without any variable modifications were considered. This translates into 168k *P. furiosus* precursors.
- For the predicted library, we converted this FASTA into a predicted *P. furiosus* spectral library using the MS²PIP Orbitrap-HCD model and our Elude RT model. Every 2+ and 3+ tryptic peptide in the proteome was predicted, with carbamidomethylation of Cysteine, and oxidation of Methionine set as respectively fixed and variable modifications. The *P. furiosus* .msp (224k precursors) was concatenated to the Human predicted .msp in EncyclopeDIA.

As decoys are generated by EncyclopeDIA, these were also appended for the *P. furiosus* proteins. All three source libraries were employed for searching the narrow window DIA data, i.e. to create a DIA-based chromatogram library. The *P. furiosus* fraction of the libraries was $\frac{5.5\text{k}}{194\text{k} + 5.5\text{k}} \approx 3\%$, $\frac{168\text{k}}{2.4\text{M} + 168\text{k}} \approx 6\%$ and $\frac{224\text{k}}{3.3\text{M} + 224\text{k}} \approx 6\%$ respectively. To account for this differential decoy fraction, the number of *P. furiosus* detections is multiplied by the inverse of their weights, using the following formula:

$$FDR = \frac{\#PyrococcusPeptides}{\#Targets} \cdot Decoyfraction\ correction$$

This corresponds to $\frac{56}{90\text{k}} \cdot \frac{194\text{k} + 5.5\text{k}}{5.5\text{k}} \approx 2\%$ for the DDA source library, $\frac{19}{46\text{k}} \cdot \frac{2.4\text{M} + 168\text{k}}{168\text{k}} \approx 1\%$ for the FASTA source library and $\frac{64}{94\text{k}} \cdot \frac{3.3\text{M} + 224\text{k}}{224\text{k}} \approx 1\%$ for the predicted source library. Note that the number of detected peptides (#targets) is slightly lower than the chromatogram libraries created without *P. furiosus* peptides (see main text). This corroborates the fact that increasing the number of false targets increases the statistical burden and thus number of false negatives, reducing the sensitivity of detection.

Prediction Models: Elude and MS²PIP

Elude: Retention Time prediction

For RT prediction, we employed Elude (version 3.02), which is available from the Percolator GitHub repository (https://github.com/percolator/percolator/releases) (2).

We trained an Elude model on the Pan-Human spectral library (4). The spectral library was downloaded from SWATHAtlas in SpectraST SPTXT file format. The peptide sequences and their respective RTs were parsed from the SPTXT file to an MS²PIP PEPREC file using the speclib_to_mgf.py script, which is available in the conversion_tools folder of the MS²PIP GitHub repository. Out of all consensus peptide spectra built from five or more identified spectra, 10000 peptides and their mean RTs were randomly sampled for training, 10000 were randomly sampled for testing and all remaining were used for final validation of the model. The training, test and validation datasets were converted and written to the Elude input file format. Through the Elude command line interface, we trained a model with the training and test subsets. Subsequently, we used the model to predict RTs for the validation subset of the dataset. The median absolute difference in experimental and predicted RTs (DeltaRT) of the validation dataset was 3.2 minutes and 95% of the DeltaRTs were less then 12.1 minutes (Figure S1). The model predictions have a Spearman rank correlation with the validation RTs of 0.98.

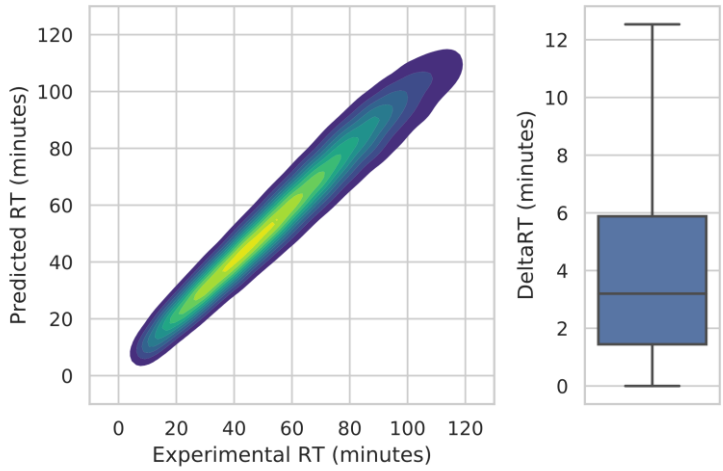

Figure S1. Evaluation of the trained Elude model. Left: Contour plot of all predicted and experimental retention times (RTs) in minutes. Right: Boxplot of all absolute differences between experimental and predicted RT (DeltaRT) in minutes. The box displays the first (Q1), second (Q2), and third (Q3) quartiles, the whiskers display Q1 - 1.5 times the interquartile range (IQR) and Q3 + 1.5 times the IQR, respectively. Outliers are not shown.

The spectral library contains oxidation and carbamidomethylation peptide modifications. As a result, the currently trained Elude model is only able to predict RTs for unmodified peptides and peptides containing these modifications. The RTs included in the original Pan-Human SPTXT spectral library are normalized to the iRT Kit peptide sequences by SpectraST. All RT values predicted by the Elude model therefore take over this normalization. As is the case for experimental RTs, the predicted RTs are aligned to the experimental dataset by EncyclopeDIA. The Elude model file is available on our GitHub repository.

MS²PIP: intensity Prediction

MS²PIP, the MS² Peak Intensity Predictor, first published by Degroeve et al., underwent significant improvements since its initial release in 2013 (12). Currently, a broad array of fragmentation models is available (e.g. Orbitrap-HCD, iontrap-CID, TripleTOF 5600+, …) (1). This gives the user the liberty to employ a model fit to the experimental setup. As both the narrow and wide window DIA datasets used in this project were obtained on a Q Exactive HF instrument (Thermo Fisher Scientific, Massachusetts, US), we employed MS²PIPs Orbitrap-HCD model, with the exception of the TT5600 model that was used for assessing PQP requirements (see main text). To further validate the application of this model, we calculated the correlations between MS²PIP predicted spectra and experimental spectra from the EncyclopeDIA DDA runs.

The Hela DDA dataset of the EncyclopeDIA article (MassIVE MSV000082805) was imported into Progenesis QI for Proteomics (Nonlinear Dynamics, Newcastle upon Tyne, UK) with default parameters. The peakpicked spectra were exported as .mgf and searched with Mascot 2.6.1 against the aforementioned human FASTA. Carbamidomethylation of Cysteine and oxidation of Methionine were respectively set as fixed and variable modifications. The precursor tolerance was set to 50 ppm and the fragment tolerance was set to 0.02 Da. The search included all 2+ and 3+ precursors, allowing up to 2 tryptic missed cleavages. Afterwards, the results were reimported into Progenesis QI for Proteomics and converted to an .msp spectral library.

The .msp spectral library was converted back to an .mgf and an MS²PIP PEPREC input format using the speclib_to_mgf.py script. Both files were then run through MS²PIP with the Orbitrap-HCD model, after which Pearson correlation coefficients (PCCs) were calculated for each experimental spectrum and its prediction. This resulted in a median PCC of 0.88 with an interquartile range of {0.795297, 0.938911} (Figure S2)

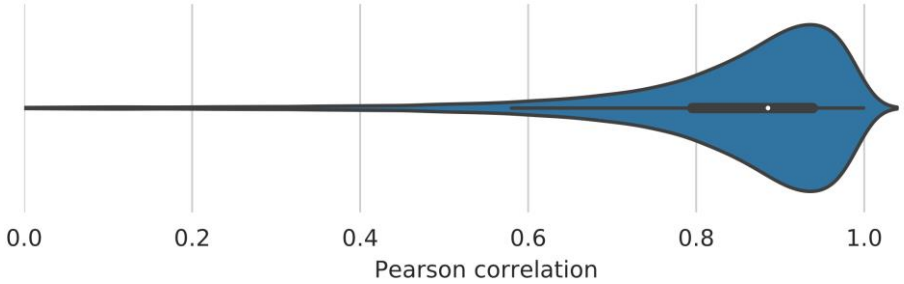

Figure S2. Violin plot showing the distribution of Pearson correlation coefficients between the MS²PIP model predictions and the experimental spectra from the EncyclopeDIA article Hela DDA dataset.

References

1. Gabriels R, Martens L, Degroeve S. Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods , instruments and labeling techniques. Nucleic Acids Res. 2019;

2. Moruz L, Staes A, Foster JM, Hatzou M, Timmerman E, Martens L, et al. Chromatographic retention time prediction for posttranslationally modified peptides. Proteomics [Internet]. 2012/05/12. 2012;12(8):1151–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22577017

3. Searle BC, Pino LK, Egertson JD, Ting YS, Lawrence RT, MacLean BX, et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun [Internet]. 2018/12/05. 2018;9(1):5128. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30510204

4. Rosenberger G, Koh CC, Guo T, Röst HL, Kouvonen P, Collins BC, et al. A repository of assays to quantify 10,000 human proteins by. Sci Data. 2014;1:1–15.

5. Ting YS, Egertson JD, Bollinger JG, Searle BC, Payne SH, Noble WS, et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods [Internet]. 2017/08/08. 2017;14(9):903–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28783153

6. Escher C, Reiter L, Maclean B, Ossola R, Herzog F, Chilton J, et al. Using iRT , a normalized retention time for more targeted measurement of peptides. Proteomics. 2012;(12):1111–21.

7. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython : freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.

8. Goloborodko AA, Levitsky LI, Ivanov M V, Gorshkov M V. Pyteomics — a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. Am Soc Mass Spectrom. 2013;(24):301–4.

9. Mallick P, Kuster B. Proteomics: a pragmatic perspective. Nat Biotechnol [Internet]. 2010/07/14. 2010;28(7):695–709. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20622844

10. Adusumilli R, Mallick P. Data Conversion with ProteoWizard msConvert. In: Comai L, Katz JE, Mallick P, editors. Proteomics: Methods and Protocols [Internet]. New York, NY: Springer New York; 2017. p. 339–68. Available from: https://doi.org/10.1007/978-1-4939-6747-6_23

11. Vaudel M, Burkhart JM, Breiter D, Zahedi RP, Sickmann A, Martens L. A complex standard for protein identification, designed by evolution. J Proteome Res [Internet]. 2012;11(10):5065–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22489649

12. Degroeve S, Martens L. MS2PIP: a tool for MS/MS peak intensity prediction. Bioinformatics [Internet]. 2013/10/01. 2013;29(24):3199–203. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24078703