Generating high-quality libraries for DIA-MS with empirically-corrected peptide predictions.

Brian C. Searle[1,2,*], Kristian E. Swearingen[1], Christopher A. Barnes[3], Tobias Schmidt[4], Siegfried Gessulat[4,5], Bernhard Kuster[4,6], and Mathias Wilhelm[4]

[1] Institute for Systems Biology, Seattle, WA, USA
[2] Proteome Software, Inc. Portland, OR, USA
[3] Novo Nordisk Research Center Seattle, Inc. Seattle, WA, USA
[4] Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany
[5] SAP SE, Potsdam, Germany
[6] Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany

*Corresponding author, email: bsearle@systemsbiology.org

Abstract

Data-independent acquisition (DIA) mass spectrometry (MS) is a powerful label-free technique for deep, proteome-wide profiling. DIA methods typically rely on sample-specific spectrum libraries from data-dependent acquisition (DDA) experiments, built at the expense of time, sample, and substantial offline fractionation effort. We report on a workflow that empirically corrects fragmentation and retention times in predicted spectrum libraries using DIA data, producing higher-quality libraries while simultaneously supplanting any need for offline fractionation or additional DDA runs.

Manuscript

Data-independent acquisition (DIA) mass spectrometry (MS) is a powerful label-free technique for deep, proteome-wide profiling[1]. DIA methods typically rely on sample-specific spectrum libraries from data-dependent acquisition (DDA) experiments, built at the expense of time, sample, and substantial offline fractionation effort. We report on a DIA-only workflow that produces higher-quality libraries while simultaneously supplanting any offline fractionation or additional DDA runs.

When sample-specific spectrum library generation is either impossible or impractical, as is frequently the case with non-model organisms, sequence variants, splice isoforms, or scarce sample quantities, software tools such as Pecan[2] and DIA-Umpire[3] can detect peptides from DIA experiments without a spectrum library by searching FASTA databases. However, these tools typically detect fewer peptides than library approaches, which leverage smaller search spaces, previously acquired instrument-specific fragmentation, and measured retention times.

Our workflow (Fig. 1a) uses a recently developed deep neural network, Prosit[4], to predict a library of fragmentation patterns and retention times for every peptide in FASTA databases. Gas-phase fractionation (GPF) improves DIA detection rates[2] by running samples multiple times with different precursor isolation windows[5]. Here we search six GPF-DIA runs of a sample pool with EncyclopeDIA[6] against the predicted library. Using these results, we empirically correct the predicted library by filtering out peptides that cannot be found in the pool, and replacing retention times and fragmentation in the predicted library with sample/instrument-specific values found in the six runs.

Although searching single-injection DIA directly with predicted libraries is highly dependent on prediction accuracy, we found that our workflow produced high-quality libraries even if the predicted libraries were not precisely tuned for the instrument, suggesting broad cross-platform applicability. We demonstrated this by modulating the normalized collision energy (NCE) setting in Prosit (Fig. 1b) and found that with a wide range of NCE settings, searching GPF-DIA of yeast on a Fusion Lumos MS produced larger empirically-corrected libraries than a sample-specific 10-fraction DDA library (Fig. 1c). Searches of single-injection DIA were less sensitive to prediction accuracy after empirical correction (Fig. 1d), which produced libraries with improved fragmentation and retention-time accuracy over both the predicted and DDA libraries (Supp. Fig. 1). Coupled with library size, these three factors had roughly equal and orthogonal improvements over directly searching single-injection DIA with

predicted libraries (Supp. Fig. 2), and, when combined, detected 66% more peptides than searching the yeast DDA library, and 37% more than a 39-injection DDA library in reanalyzed HeLa datasets[6] (Fig. 1e).

We applied our workflow to analyze gametocyte cultures of *Plasmodium falciparum*, the parasite responsible for 50% of all malaria cases. Using a QE-HF MS, we produced a 3,378 protein empirically-corrected library, enabling measurement of 2,909 proteins in single-injection DIA experiments, and increasing the previously known proteome[7] size by 58% (Supp. Fig. 3). With this library, we detected parasite proteins in up to 1:99 dilution with uninfected red blood cells in single-injection DIA experiments (Fig. 1f), while maintaining higher quantification accuracy and depth than comparable DDA experiments (Fig. 1g, Supp. Fig. 4).

In conclusion, empirical correction of predicted spectrum libraries enables rapid library generation for non-canonical proteomes or non-model organisms without offline fractionation.

Data availability statement. Prosit (https://www.proteomicsdb.org/prosit) and EncyclopeDIA 1.0 (https://bitbucket.org/searleb/encyclopedia) are both available under the Apache 2 open-source license. The raw data from the yeast and *P. falciparum* studies are available at MassIVE (MSV000084000). The raw data from the HeLa reanalysis are available as originally published at MassIVE (MSV000082805). We also provide a graphical workflow (Supplementary Note) to facilitate making empirically-corrected libraries for new proteomes from FASTA databases.

Acknowledgements. We thank S. Kappe for insightful discussions, N. Carmago for providing the gametocyte samples, L. Pino for providing yeast lysates, B. Kim for technical assistance, as well as M. MacCoss, K. Grove, and M. Guldbrandt for providing instrument time. B.C.S. is supported by the Translational Research Fellows Program (TRFP) from the Institute for Systems Biology. K.E.S. is supported by K25 AI119229. This work was in part funded by the

German Federal Ministry of Education and Research (BMBF, grant no. 031L0008A and no. 031L0168) and the EU Horizon 2020 grant EPIC-XS (grant no. 823839).

Author contributions. B.C.S. conceived the study. B.C.S., K.E.S., and C.A.B. performed the experiments. B.C.S., T.S., and S.G. developed the software. B.C.S., B.K., and M.W. supervised the work. All authors wrote and approved the manuscript.

Competing financial interests Statement. B.C.S. is a founder and shareholder in Proteome Software, which operates in the field of proteomics. M.W. and B.K. are founders and shareholders of OmicScouts, but have no operational role in the company.

References:

1. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11**, O111.016717 (2012).

2. Ting, Y. S. *et al.* PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017).

3. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258 (2015).

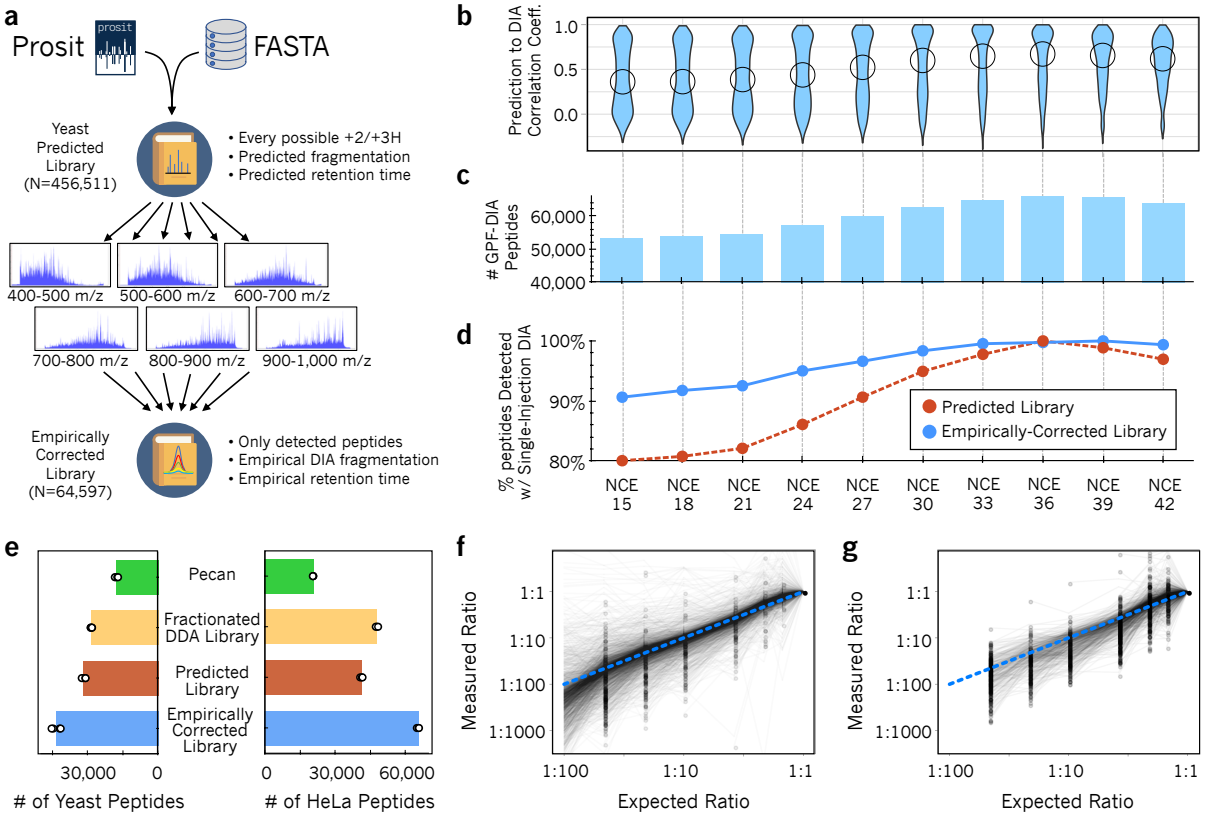
4. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).

5. Panchaud, A. *et al.* Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81**, 6481–6488 (2009).

6. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).

7. Lasonder, E. *et al.* Integrated transcriptomic and proteomic analyses of P. falciparum gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res.* **44**, 6087–6101 (2016).

Figure 1: Evaluating empirically-corrected libraries made with peptide predictions. (**a**) Workflow for generating empirically-corrected libraries from Prosit predictions of peptides in FASTA databases. GPF-DIA runs of a sample pool are searched with a spectrum library of fragmentation and retention time predictions. Peptides detected in these runs are compiled with fragmentation patterns and retention times extracted from the GPF-DIA data into an empirically-corrected library. (**b**) Violin plots showing spectral correlation between predicted library for yeast peptides and single-injection DIA at various NCE settings (circles indicate medians). (**c**) The total numbers of empirically-corrected library entries detected from GPF-DIA at various NCE settings. (**d**) The fraction of peptides detected in single-injection DIA relative to the optimal NCE for empirically-corrected libraries (blue line) are more consistent than predicted libraries (red dashed line) across a wide range of NCE settings. (**e**) The average number of yeast (N=4) and HeLa (N=3) peptides in single-injection DIA detected using Pecan, a fractionated spectrum library, a predicted spectrum library, and an empirically-corrected library. Quantitative ratios of (**f**) 2,740 DIA- and (**g**) 1,491 DDA-detected malaria parasite *P. falciparum* proteins (gray lines) after dilution with uninfected RBC lysate relative to the undiluted sample. Dots indicate the lowest dilution level of measurement for each protein, and the blue dashed line indicates the ideal measurement. No *P. falciparum* proteins were detected or quantified in the last dilution sample (1:99) using DDA.


a
Prosit
FASTA
Yeast Predicted Library (N=456,511)
• Every possible +2/+3H
• Predicted fragmentation
• Predicted retention time
400-500 m/z
500-600 m/z
600-700 m/z
700-800 m/z
800-900 m/z
900-1,000 m/z
Empirically Corrected Library (N=64,597)
• Only detected peptides
• Empirical DIA fragmentation
• Empirical retention time
b
Prediction to DIA Correlation Coeff.
1.0
0.5
0.0
c
GPF-DIA Peptides
60,000
50,000
40,000
d
% peptides Detected w/ Single-Injection DIA
100%
90%
80%
Predicted Library
Empirically-Corrected Library
NCE 15
NCE 18
NCE 21
NCE 24
NCE 27
NCE 30
NCE 33
NCE 36
NCE 39
NCE 42
e
Pecan
Fractionated DDA Library
Predicted Library
Empirically Corrected Library
30,000
0
# of Yeast Peptides
0
30,000
60,000
# of HeLa Peptides
f
Measured Ratio
1:1
1:10
1:100
1:1000
1:100
1:10
1:1
Expected Ratio
g
Measured Ratio
1:1
1:10
1:100
1:1000
1:100
1:10
1:1
Expected Ratio

1

Supplementary Methods:

***Saccharomyces cerevisiae* culture and sample preparation.** As described in Searle *et al*. [6], *S. cerevisiae* strain BY4741 (Dharmacon) was cultured at 30°C in YEPD and harvested at mid-log phase. Cells were pelleted and lysed in a buffer of 8 M urea, 50 mM Tris (pH 8), 75 mM NaCl, 1 mM EDTA (pH 8) followed by 7 cycles of 4 minutes bead beating with glass beads. After a 1 minute rest on ice, lysate was collected by piercing the tube and centrifuging for one minute at 3,000 x g and 4°C into an empty eppendorf. After further centrifugation at 21,000 x g and 4°C for 15 minutes, the protein content of the supernatant was removed and estimated using BCA. Proteins were then reduced with 5 mM dithiothreitol for 30 minutes at 55°C, alkylated with 10 mM iodoacetamide in the dark for 30 minutes at room temperature, and diluted to 1.8 M urea, before digestion with sequencing grade trypsin (Pierce) at a 1:50 enzyme to substrate ratio for 16 hours at 37°C. 5N HCl was added to approximately pH 2 to quench the digestion, and the resulting peptides were desalted with 30 mg MCX cartridges (Waters). Peptides were dried with vacuum centrifugation and brought to 1 µg / 3 µl in 0.1% formic acid (buffer A) prior to mass spectrometry acquisition.

***Plasmodium falciparum* culture and red blood cell sample preparation.** Three flasks of stage IV/V *P. falciparum* NF54 gametocytes were prepared. Asexual cultures were synchronized with sorbitol and set up at 5% hematocrit and 1% parasitemia. Gametocytogenesis was induced by withholding fresh blood and allowing parasitemia to increase. N-acetyl glucosamine was added to media for 4 days beginning 7 days after set-up in order to remove asexual parasites. Gametocyte-infected erythrocytes (giRBC) were enriched from uninfected erythrocytes (uiRBC) by magnetic-activated cell sorting at stage III. Stage IV/V gametocytes were collected on day 15 post-set-up. Additional uiRBC were also prepared by washing multiple times in RPMI and stored at 50% hematocrit. giRBC and uiRBC cells were lysed in a buffer of 10% sodium dodecyl sulfate (SDS), 100 mM ammonium bicarbonate (ABC),

cOmplete EDTA-free Protease Inhibitor Cocktail (Sigma), and Halt Phosphatase Inhibitor Cocktail (Thermo Scientific). Proteins were then reduced with 20-40 mM tris(2-carboxyethyl)phosphine (TCEP) for 10 minutes at 95°C and alkylated with 40-80 mM iodoacetamide in the dark for 20 minutes at room temperature. After centrifugation at 16,000 x g to pellet insoluble material, proteins were purified with methanol:chloroform extraction[8] and dried and resuspended in 8M urea buffer before the content was estimated using BCA. After dilution to 1.8 M urea, proteins were digested with sequencing-grade trypsin (Promega) at a 1:40 enzyme-to-substrate ratio for 15 hours at 37°C. The resulting peptides were desalted with Sep-Pak cartridges (Waters), dried with vacuum centrifugation, and brought to 1 µg / 3 µl in 0.1% formic acid (buffer A) prior to mass spectrometry acquisition. In addition, several digested peptide mixtures were made by diluting peptides from one flask of giRBC cells with peptides from uiRBC cells at ratios of 1:0, 2:1, 7:8, 4:15, 1:9, 2:41, 2:91, and 1:99 giRBC:uiRBC.

Liquid chromatography mass spectrometry (*S. cerevisiae*). Tryptic *S. cerevisiae* peptides were separated with a Thermo Easy nLC 1200 on self-packed 30 cm columns packed with 1.8 µm ReproSil-Pur C18 silica beads (Dr. Maisch) inside of a 75 µm inner diameter fused silica capillary (#PF360 Self-Pack PicoFrit, New Objective). The 30 cm column was coiled inside of a Sonation PRSO-V1 column oven set to 35°C prior to ionization into the MS. The HPLC was performed using 200 nl/min flow with solvent A as 0.1% formic acid in water and solvent B as 0.1% formic acid in 80% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted with a linear gradient from 7% to 38% buffer B over 90 minutes. Following the linear separation, the system was ramped up to 75% buffer B over 5 minutes and finally set to 100% buffer B for 15 minutes, which was followed by re-equilibration to 2% buffer B prior to the subsequent injection. Data were acquired using data-independent acquisition (DIA).

The Thermo Fusion Lumos was set to acquire 6 GPF-DIA acquisitions of a biological sample pool using 120,000 precursor resolution and 30,000 fragment resolution. The automatic

gain control (AGC) target was set to 4e5, the maximum ion inject time (IIT) was set to 60 ms, the normalized collision energy (NCE) was set to 33, and +2H was assumed as the default charge state. The GPF-DIA acquisitions used 4 m/z precursor isolation windows in a staggered window pattern with optimized window placements (i.e., 398.4 to 502.5 m/z, 498.5 to 602.5 m/z, 598.5 to 702.6 m/z, 698.6 to 802.6 m/z, 798.6 to 902.7 m/z, and 898.7 to 1002.7 m/z). Individual samples for proteome profiling acquisitions used single-injection DIA acquisitions (120,000 precursor resolution, 15,000 fragment resolution, AGC target of 4e5, max IIT of 20 ms) using 8 m/z precursor isolation windows in a staggered window pattern with optimized window placements from 396.4 to 1004.7 m/z.

For generation of an *S. cerevisiae* spectral library, 80 µg of the same tryptic digests described above were separated into 10 total fractions using the Pierce high-pH reversed-phase peptide fractionation Kit (Thermo, #84868). Briefly, peptides were loaded onto hydrophobic resin spin column and eluted using the following 8 acetonitrile steps: 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20.0%, and 50%, keeping both the wash and flow-through. The resulting peptide fractions were injected into the same Thermo Fusion Lumos using the same chromatography setup and column described above, but configured for data-dependent acquisition (DDA). After adjusting each fraction to an estimated 0.5 to 1.0 µg on column, the fractions were measured in a top-20 configuration with 30 second dynamic exclusion. Precursor spectra were collected from 300-1650 m/z at 120,000 resolution (AGC target of 4e5, max IIT of 50 ms). MS/MS were collected on +2H to +5H precursors achieving a minimum AGC of 2e3. MS/MS scans were collected at 30,000 resolution (AGC target of 1e5, max IIT of 50 ms) with an isolation width of 1.4 m/z with a NCE of 33.

Liquid chromatography mass spectrometry (*P. falciparum*). Tryptic *P. falciparum* and RBC peptides were separated with a Thermo Easy nLC 1000 and emitted into a Thermo Q-Exactive HF. In-house laser-pulled tip columns were created from 75 µm inner diameter fused silica

capillary and packed with 3 µm ReproSil-Pur C18 beads (Dr. Maisch) to 30 cm. Trap columns were created from Kasil fritted 150 µm inner diameter fused silica capillary and packed with the same C18 beads to 2 cm. The HPLC was performed using 250 nl/min flow with solvent A as 0.1% formic acid in water and solvent B as 0.1% formic acid in 80% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted using a 84-minute gradient from 6 to 40% buffer B, followed by steep 5-minute gradient from 40 to 75% buffer B and finally set to 100% buffer B for 15 minutes, which was followed by re-equilibration to 0% buffer B prior to the subsequent injection. Data were acquired using either DDA or DIA.

The Thermo Q-Exactive HF was set to acquire DDA in a top-20 configuration with “auto” dynamic exclusion. Precursor spectra were collected from 400-1600 m/z at 60,000 resolution (AGC target of 3e6, max IIT of 50 ms). MS/MS were collected on +2H to +5H precursors achieving a minimum AGC of 1e4. MS/MS scans were collected at 15,000 resolution (AGC target of 1e5, max IIT of 25 ms) with an isolation width of 1.4 m/z with a NCE of 27. Additionally, 6 GPF-DIA acquisitions were acquired of a biological sample pool (60,000 precursor resolution, 30,000 fragment resolution, AGC target of 1e6, max IIT of 60 ms, NCE of 27, +3H assumed charge state) using 4 m/z precursor isolation windows in a staggered window pattern with optimized window placements (i.e., 398.4 to 502.5 m/z, 498.5 to 602.5 m/z, 598.5 to 702.6 m/z, 698.6 to 802.6 m/z, 798.6 to 902.7 m/z, and 898.7 to 1002.7 m/z). Individual samples used single-injection DIA acquisitions (60,000 precursor resolution, 30,000 fragment resolution, AGC target of 1e6, max IIT of 60 ms) using 16 m/z precursor isolation windows in a staggered window pattern with optimized window placements from 392.4 to 1008.7 m/z.

FASTA databases and predicted spectrum libraries. Species-specific reviewed FASTA databases for *Homo sapiens* (April 25 2019, 20415 entries) and *Saccharomyces cerevisiae* (January 25 2019, 6729 entries) were downloaded from Uniprot. The *Plasmodium falciparum*[9] FASTA database was downloaded from PlasmoDB[10] version 43 (April 24 2019, 5548 entries).

Each database was digested *in silico* to create all possible +2H and +3H peptides with precursor m/z within 396.43 and 1002.70, assuming up to one missed cleavage. Peptides were further limited to be between 7 and 30 amino acids to match the restrictions of the Prosit tool[4]. In general, NCE were assumed to be 33 (yeast was processed using NCE from 15 to 42 in 3 NCE increments) but modified to account for charge state. Since DIA assumes all peptides are of a fixed charge, we adjusted the NCE setting as if peptides were fragmented at the wrong charge state using the formula:

$$Adjusted\ NCE = NCE * \frac{factor(default\ charge)}{factor(peptide\ charge)}$$

where the factors were 1.0 for +1H, 0.9 for +2H, 0.85 for +3H, 0.8 for +4H, and 0.75 for +5H and above[11]. After submitting to Prosit, predicted MS/MS and retention times were converted to the EncyclopeDIA DLIB format for further processing. Scripts to produce Prosit input from FASTAs and build EncyclopeDIA-compatible spectrum libraries from Prosit output are available as functions in EncyclopeDIA 1.0.

DDA data processing. All Thermo RAW files were converted to .mzML format using the ProteoWizard package[12] (version 3.0.18299) using vendor peak picking. DDA data was searched with Comet[13] (version 2017.01 rev. 1), allowing for fixed cysteine carbamidomethylation, variable peptide n-terminal pyro-glu, and variable protein n-terminal acetylation. Fully tryptic searches were performed with a 50 ppm precursor tolerance and a 0.02 Da fragment tolerance permitting up to 2 missed cleavages. High-pH reverse-phase fractions were combined and search results were filtered to a 1% peptide-level FDR using PeptideProphet[14] from the Trans-Proteomic Pipeline[15] (TPP version 5.1.0). A yeast-specific Bibliospec[16] DDA spectrum library was created from Thermo Q-Exactive DDA data using Skyline[17] (Daily version 19.0.9.149).

P. falciparum and RBC DDA data was additionally processed with MaxQuant[18] (version 1.6.5.0) to perform label-free quantitation with precursor ion integration. MaxQuant was configured to use default parameters, briefly fixed cysteine carbamidomethylation, variable methionine oxidation, and variable protein n-terminal acetylation. Fully tryptic searches were performed with a 20 ppm fragment tolerance using both the human and *P. falciparum* FASTA databases, as well as common contaminants and filtered to a 1% peptide-level FDR. Quantification was performed using unique and razor peptides with the match between runs setting turned on.

DIA data processing. DIA data was overlap demultiplexed[19] with 10 ppm accuracy after peak picking in ProteoWizard (version 3.0.18299). Searches were performed using EncyclopeDIA (version 0.8.3), which was configured to use default settings: 10 ppm precursor, fragment, and library tolerances. EncyclopeDIA was allowed to consider both B and Y ions and trypsin digestion was assumed. EncyclopeDIA search results were quantified and filtered to a 1% peptide-level using Percolator[20] (version 3.1) and 1% protein-level FDR assuming parsimony.

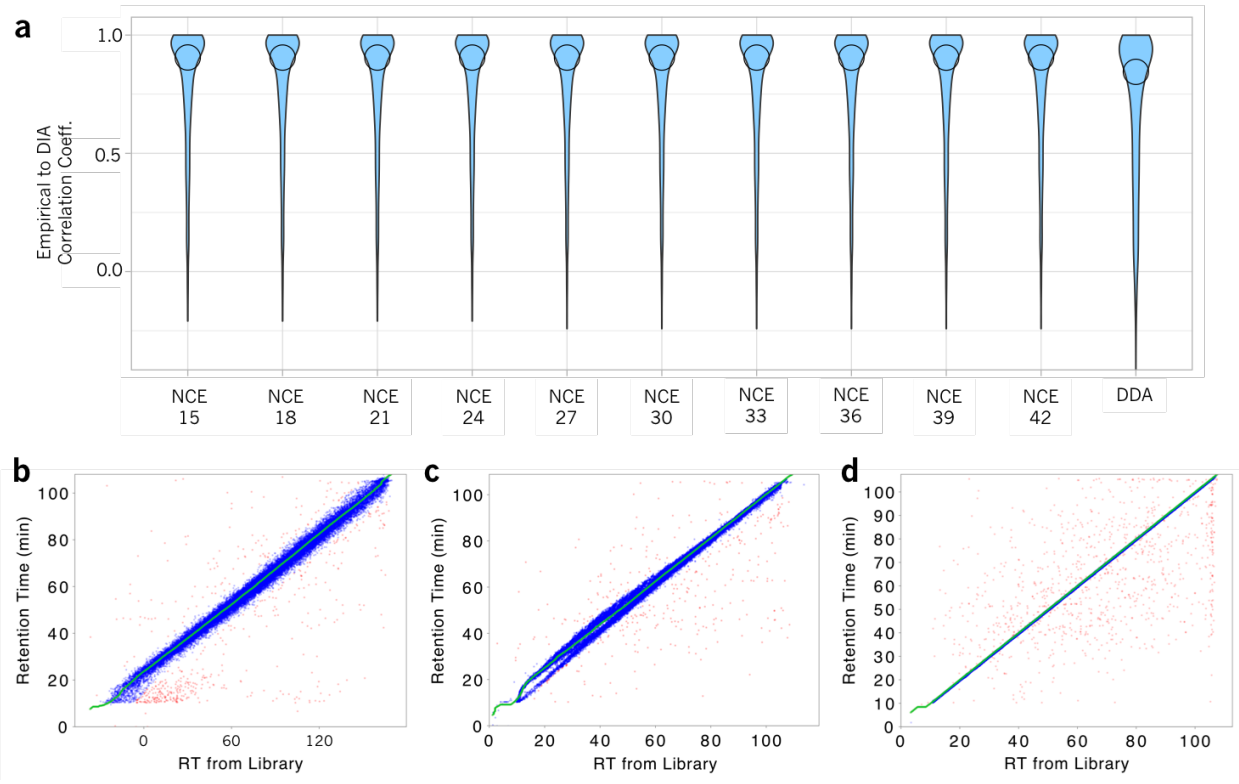
Empirically-corrected library generation. Predicted libraries were corrected with EncyclopeDIA using the chromatogram library method described previously[6]. Briefly, GPF-DIA samples for a given study were loaded into EncyclopeDIA using the above parameters, where the search library was set to the appropriate predicted spectrum library. Peptides detected by EncyclopeDIA were exported as a chromatogram library.

A peptide entry in a chromatogram library is similar to a peptide entry in a spectrum library in that it contains a precursor mass, retention time, and a fragmentation spectrum. In addition, a chromatogram library entry also contains the peptide peak shape, and a correlation score for every fragment ion indicating the agreement between the overall peptide peak shape and the fragment peak shape. This score provides an indication of the likelihood the fragment

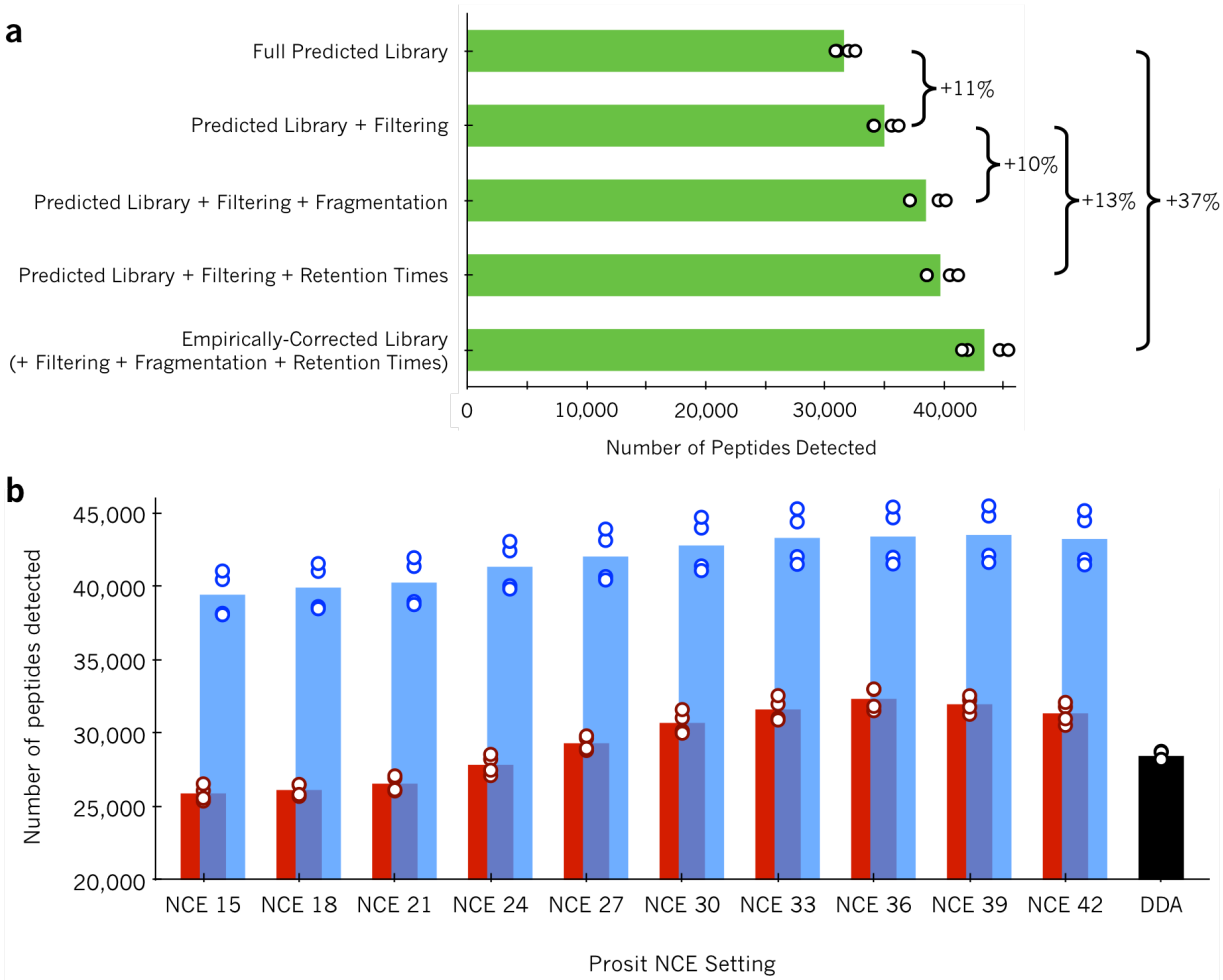
ion was interfered with in the GPF-DIA, under the expectation that it will also likely be interfered with in single-injection DIA as well.

This process created a new empirically-corrected library containing only peptides found in the GPF-DIA samples, and also retained empirical fragment ion intensities and retention times observed from the DIA data. These libraries were made to be compatible with both EncyclopeDIA and Skyline and were used for downstream analysis of single-injection DIA.

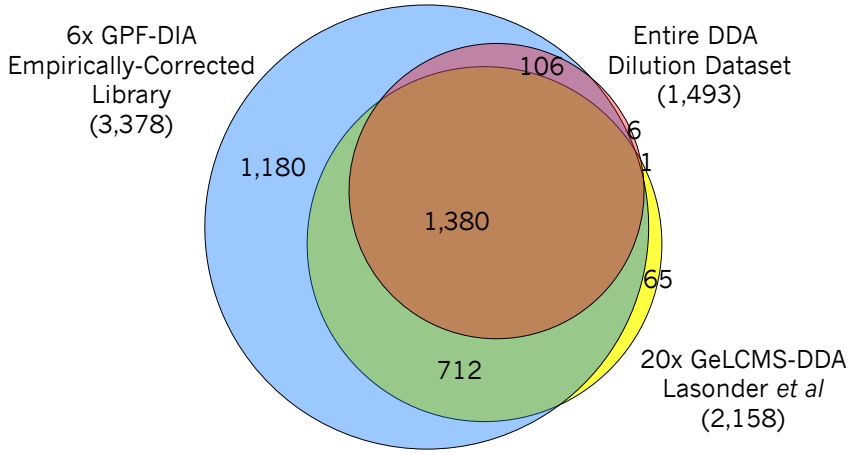
Supplementary Figures:



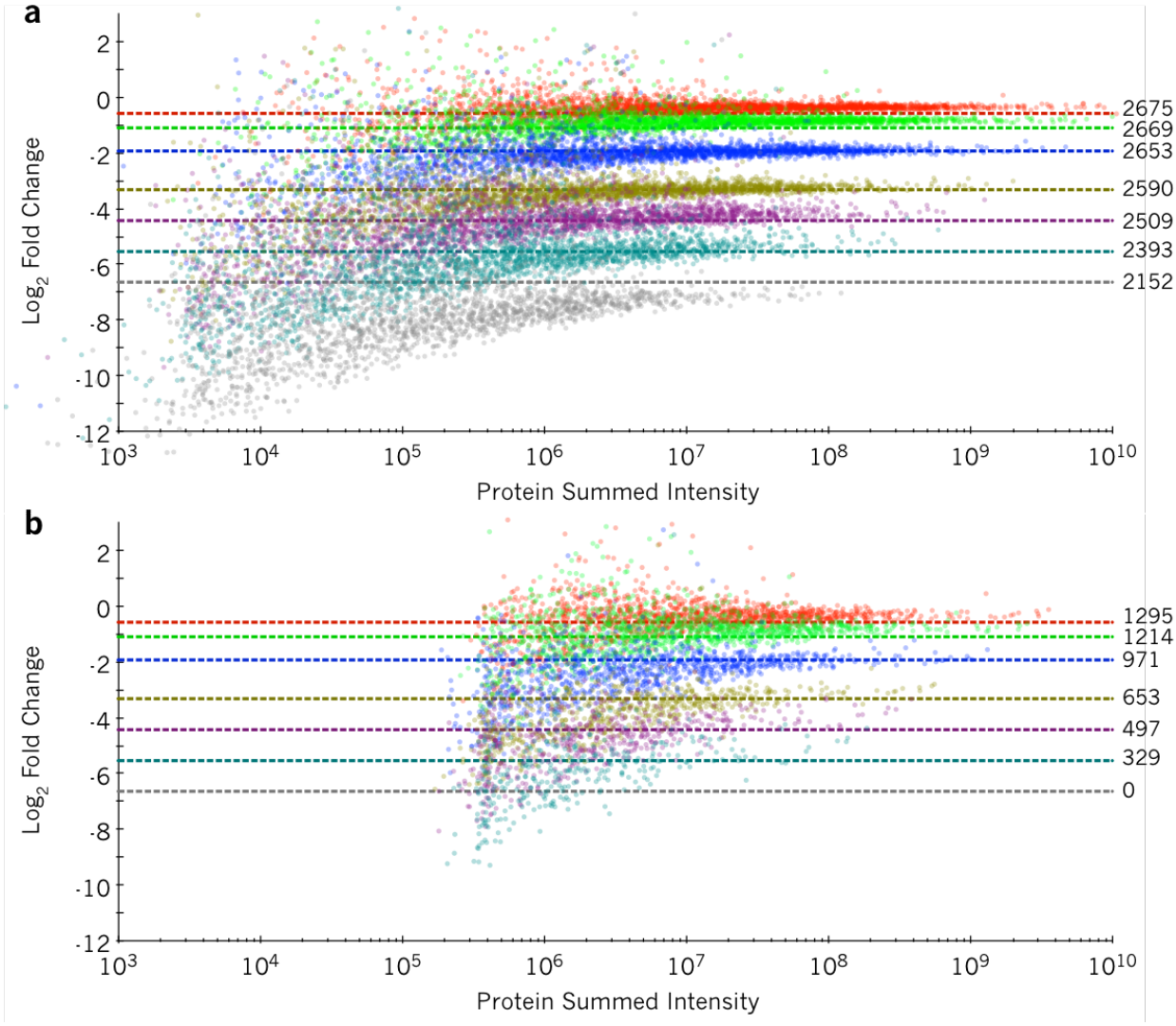
Supplementary Figure 1: **Comparison of empirically-corrected library and DDA library fragmentation and retention times.** (**a**) Violin plots show that after empirical correction, the correlation between empirically-corrected library fragmentation and the integrated DIA chromatograms is consistently higher than the high-pH reverse-phase fractionated DDA spectrum library. Retention time alignment between one single-injection yeast replicate and (**b**) the DDA library, (**c**) the Prosit predicted library, and (**d**) the resulting empirically-corrected library. Even though the DDA library was acquired on the same instrument with the same chromatography setup, there is error (including some minor streaking) between the fractionated and the single-injection retention times due to matrix effects. The empirically-corrected libraries have improved fragmentation and retention time accuracy over the predicted library and the DDA library, as GPF-DIA fragmentation patterns closely resemble single-injection DIA, and GPF does not affect peptide interactions with matrix.



Supplementary Figure 2: **Additive effect of library size filtering, retention time, and fragmentation patterns.** (**a**) The number of yeast peptides detected at 1% peptide FDR in single-injection DIA runs (N=4) using the NCE=33 chromatogram library, where either the retention times or fragmentation patterns have been switched with the predicted Prosit values. Compared to the predicted spectrum library search, an 11% increase comes from simply using a narrowed peptide selection. DIA-based retention times and fragmentation patterns provide a 13% and 10% increase over this, respectively. Comparing the chromatogram library and the predicted library detections (37% increase), these percentage gains appear to be nearly multiplicative (i.e., 111%*110%*113%=138%), indicating that all 3 factors are independent and of roughly equal importance. (**b**) The same single-injection DIA runs (N=4) searched against either the Prosit predicted library (red bars) at various NCE settings, the resulting chromatogram libraries (blue bars), or a high-pH reverse-phase fractionated DDA spectrum library (black bar).



Supplementary Figure 3: **The *P. falciparum* chromatogram library is a superset of previously reported proteomes.** A proportionally-sized Euler diagram showing the overlap between proteins detected in the chromatogram library (NCE=33), the full single-injection dilution curve with DDA, and the Stage IV/V mixed male/female sexual *P. falciparum* proteome published by Lasonder *et al.*[7] from 20x GeLCMS-fractionated DDA runs.



Supplementary Figure 4: **Quantification of *P. falciparum* proteins.** Quantitative ratios using either (**a**) DIA or (**b**) DDA for *P. falciparum* proteins at 7 different dilution ratios with red blood cell lysates (red=2:1, green=7:8, blue=4:15, gold=1:9, purple=2:41, cyan=2:91, and gray=1:99) relative to the protein intensity (summed from peptide intensities). Dashed lines indicate the expected ratio, where the number of proteins in each measurement batch are indicated at the right. No *P. falciparum* proteins were detected or quantified in the 1:99 dilution sample using DDA.

Supplementary Note:
Using EncyclopeDIA and Prosit to Make Empirically-Corrected Peptide Libraries

Data-independent acquisition (DIA) mass spectrometry is a powerful label-free proteomics technique. DIA methods typically rely on sample-specific spectrum libraries from deeply fractionated data-dependent acquisition (DDA) experiments. This note (**based on EncyclopeDIA version 0.8.3**) will walk you through how to create equal- or higher-quality libraries using only DIA data.

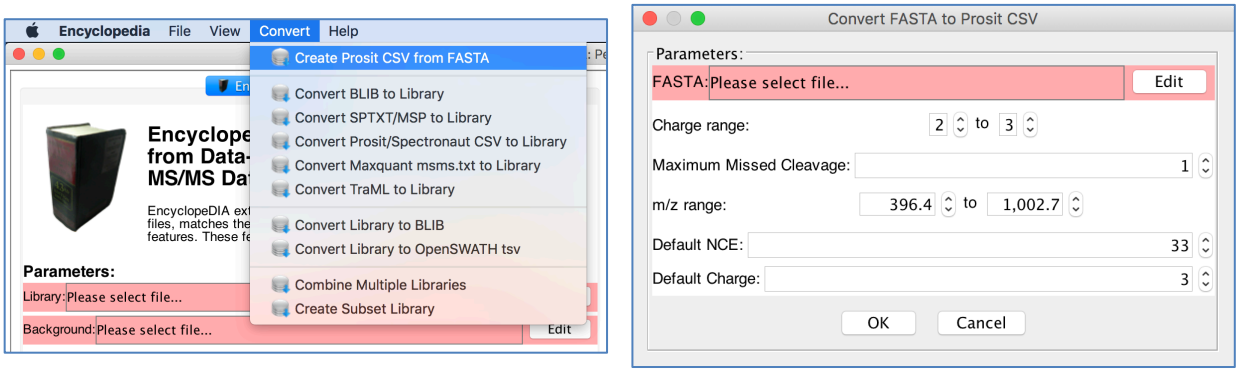
Prerequisites:
Prosit is a web application that is accessible from any operating system without installation. EncyclopeDIA is a cross-platform Java application that has been tested for Windows, Macintosh, and Linux. EncyclopeDIA requires 64-bit Java 1.8 or higher. If you don't already have it, you can download "Windows x64" from:
http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html

Alternatively, you can also use the open-source OpenJDK if the Oracle's Java license is too restrictive. After you have 64-bit Java 1.8, double click on the EncyclopeDIA .JAR file to launch the GUI interface. If you are using a Macintosh, you may need to right click on the EncyclopeDIA .JAR and select "Open" to execute it for the first time with the proper permissions.

Collecting DIA Data:
We recommend consulting this general-purpose overview for collecting DIA data:
https://bitbucket.org/searleb/encyclopedia/downloads/dia%20methods%20setup.pdf

We recommend these parameters for collecting single-injection and gas-phase fractionated DIA:
https://docs.google.com/spreadsheets/d/1A8AQlmLroAkQcAcsiGTNvnGBE2lGpkMwhh0YLTBHXKA

Building Prosit CSV Input with EncyclopeDIA:
The "Convert/Create Prosit CSV from FASTA" menu option launches a dialog for building a Prosit CSV. Select the input FASTA, charge range, the number of maximum missed cleavages, and the m/z range of interest. In general, because FDR is calculated based on the total number of peptides searched, we recommend using small FASTA databases (fewer than 25,000 entries) and narrow charge ranges to shrink the necessary search space:



Based on how you acquired your DIA experiment, set the target NCE and default charge state for the prediction. With DIA all peptides are fragmented assuming they are the same charge. For peptides that are not the default charge state, EncyclopeDIA back-calculates what the Prosit NCE should be. NCE settings are based on the Thermo Fusion Lumos mass spectrometer. The default (33 NCE) should work in most cases, but if you'd like to fine-tune this setting then the following conversions are a good starting point:

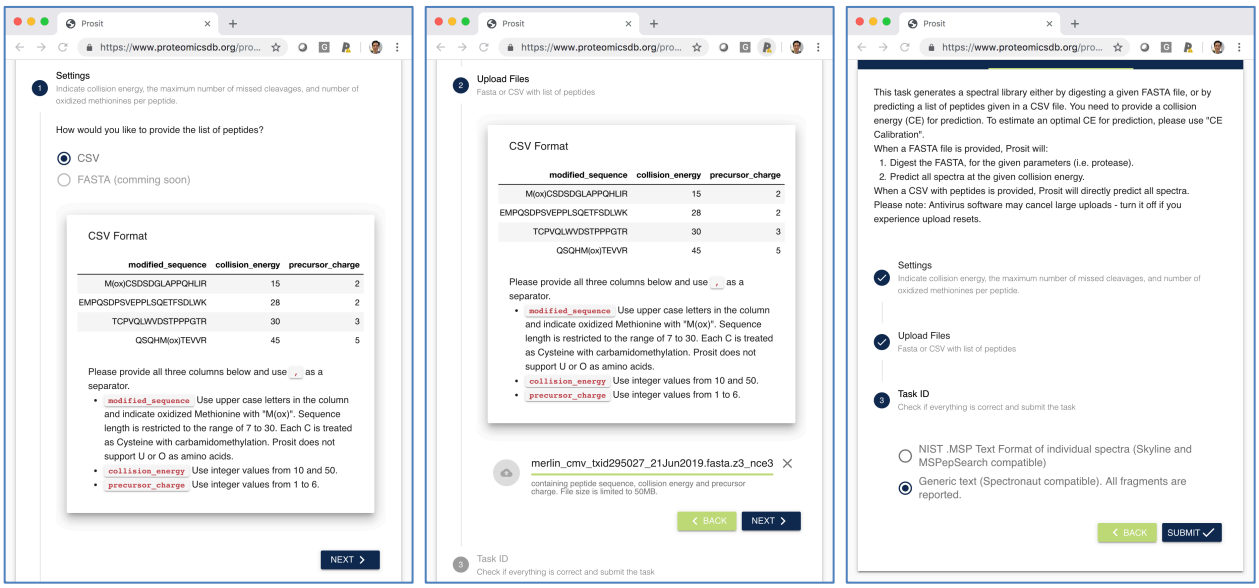
- Fusion-class orbitrap instruments: Actual NCE setting
- Q-Exactive-class orbitrap instruments: Actual NCE setting + 6
- ToF instruments: EV/2.5

This will create a CSV file in the same directory as your FASTA.

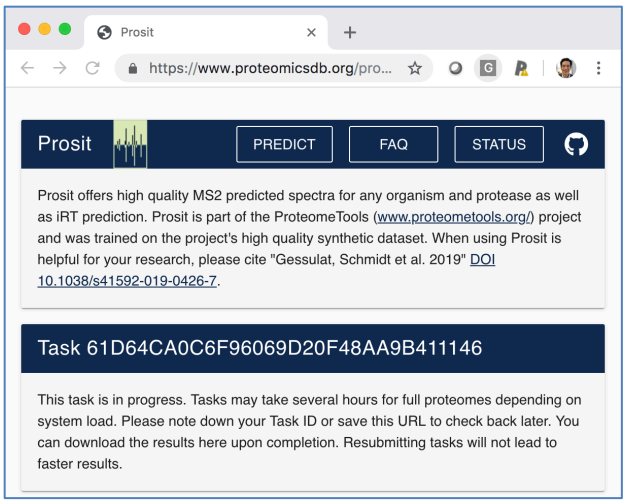
Generating Prosit Predictions:

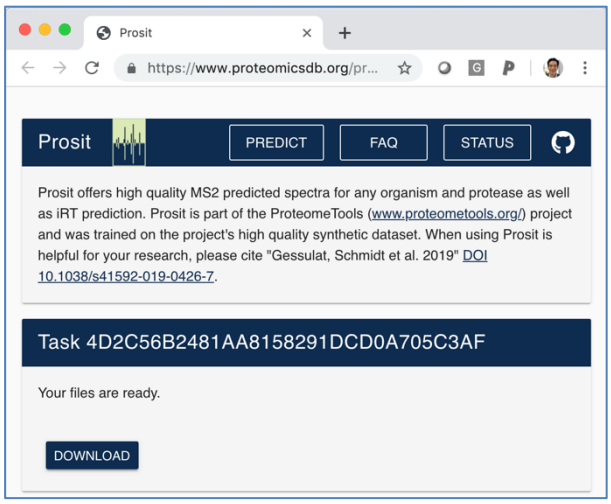
First, go to the Prosit website at https://www.proteomicsdb.org/prosit/. Using the "Spectral Library" tab, follow these three steps:

1. "Settings": select "CSV" to provide the list of peptides, then hit "next"
2. "Upload Files": click the cloud icon to upload the CSV from EncyclopeDIA, then hit "next"
3. "Task ID": select "Generic text" for the return format and hit "submit"

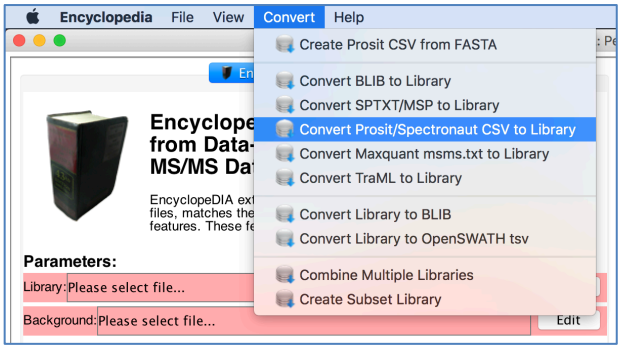


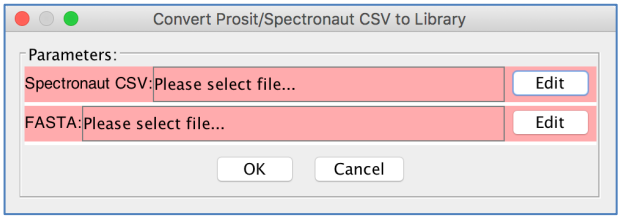
After the job has been submitted, record your Task ID. Depending on the size of the FASTA, your task may take hours to days to process. You can refresh the URL to check if your job has finished, or save the URL to check back at a later time. Once your task is complete, download the resulting files:





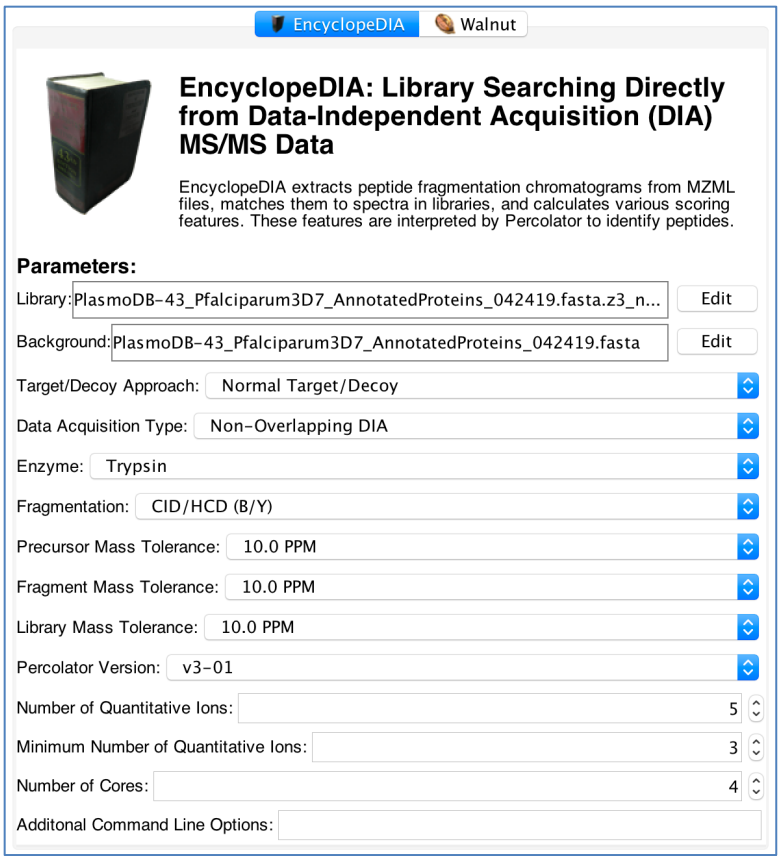
Creating an EncyclopeDIA Library from Prosit CSV Output:
The “Convert/Create Prosit CSV from FASTA” menu option launches a dialog for building a predicted library. Upload the Prosit output CSV and your original FASTA database:



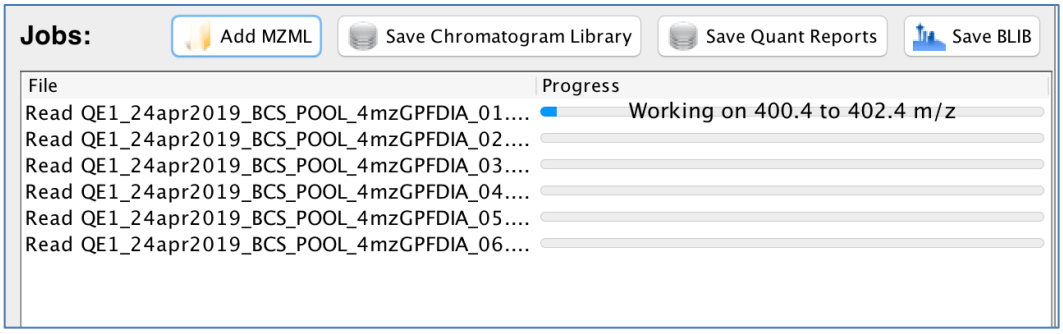


This will create a DLIB library in the same directory as your CSV.

Making an Empirically-Corrected Library from Gas-Phase Fractionated Runs:
Using EncyclopeDIA parameter tab, specify the DLIB library and your FASTA database. Then set up your EncyclopeDIA search using settings appropriate for your experiment. For example, the following settings are appropriate for most orbitrap mass spectrometers:



Then, queue up mzML files from your gas-phase fractionated runs. We recommend using MSConvert (Proteowizard) for building vendor-neutral mzML files from your vendor-specific raw files:



Once your files have finished processing, press the button “Save Chromatogram Library” to create your empirically-corrected library in an ELIB format. You can use this ELIB library in EncyclopeDIA, Skyline, or Scaffold DIA to analyze your single-injection DIA experiments.

Supplementary References:

8. Wessel, D. & Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138**, 141–143 (1984).

9. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).

10. Aurrecoechea, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–43 (2009).

11. Orsburn, B. Normalized collision energy calculation for Q Exactive. (2014). Available at: http://proteomicsnews.blogspot.com/2014/06/normalized-collision-energy-calculation.html. (Accessed: 13th June 2019)

12. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

13. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).

14. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).

15. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.* **9**, 745–754 (2015).

16. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**, 5678–5684 (2006).

17. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

18. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

19. Amodei, D. *et al.* Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J. Am. Soc. Mass Spectrom.* (2019). doi:10.1007/s13361-018-2122-8

20. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).