Picky – a simple online method designer for targeted proteomics

Henrik Zauber[1] and Matthias Selbach[1]

1 Proteome Dynamics, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, D-13092 Berlin, Germany

Corresponding author:

Matthias Selbach
Tel.: +49 30 9406 3574
Fax.: +49 30 9406 2394
email: matthias.selbach@mdc-berlin.de

Targeted proteomic approaches like selected reaction monitoring (SRM) and parallel reaction monitoring (PRM) are increasingly popular because they enable sensitive and rapid analysis of a preselected set of proteins[1-3]. However, developing targeted assays is tedious and requires the selection, synthesis and mass spectrometric analysis of candidate peptides before the actual samples can be analyzed. The SRMatlas and ProteomeTools projects recently published fragmentation spectra of synthetic peptides covering the entire human proteome[4,5]. These datasets provide very valuable resources. However, extracting the relevant data for selected proteins of interest is not straightforward. For example, developing scheduled acquisition methods (i.e. analyzing specific peptides in defined elution time windows) is complicated and requires adjustments to specific chromatographic conditions employed. Moreover, the number of peptide candidates to be targeted in parallel often exceeds the analytical abilities of the mass spectrometer. In this case, the key question is which peptides can be omitted without losing too much information. Ideally, a method design tool would automatically select the most informative peptides in each retention time window. Until now, none of the available tools automatically generates such optimized scheduled SRM and PRM methods (Figure S1).

Here, we present Picky (https://picky.mdc-berlin.de): a fast and easy to use online tool to design scheduled PRM/SRM assays (Figure 1). Users only need to provide identifiers for

human proteins of interest. Based on this input, Picky selects corresponding tryptic peptides from the ProteomeTools dataset for targeted analysis. Picky comes with a scheduling algorithm that adapts to different HPLC gradients (see Figure S2). To this end, users can provide a list of experimentally observed peptide retention times on their HPLC system. A simple shotgun analysis of any standard sample will generate such a list. Picky uses these data to estimate retention times of peptides to be targeted via their hydrophobicity scores. Importantly, the resulting acquisition list is further optimized if the number of peptides monitored in parallel exceeds a user defined threshold. In this case, the lowest scoring peptide from the protein with the highest number of targeted peptides is removed. This step is repeated until the number of peptides to be targeted in parallel has reached the desired threshold. Hence, Picky selects the best set of peptides covering the proteins of interest under the constraints of the HPLC gradient employed. Parameters such as fragmentation types, charge states and retention time windows can be adjusted dynamically. For SRM, Picky selects transitions based on the most intense fragment ions observed. The tool exports an inclusion list, which can be imported into the acquisition software of the mass spectrometer. In addition, Picky displays and exports annotated fragmentation spectra and a spectral library for all targeted peptides. This library can be imported into Skyline[6] to validate the acquired SRM/PRM data via intensity correlation methods.

To assess the performance of PRM methods designed by Picky we carried out a benchmark experiment. As reference samples we used different amounts of human proteins spiked into 1.4 µg yeast lysate. We only provided Picky with identifiers of proteins to be targeted and a list of experimentally observed peptide retention times. Based on this input, Picky designed an optimized PRM method in less than a minute. We then used this method to analyse the reference samples by PRM. For comparison, we also analyzed the same samples via standard data dependent acquisition (DDA). PRM markedly outperformed DDA at higher dilutions of the spiked-in proteins (Figure S3). For example, at 300 attomoles PRM still identified 31 of the 45 targeted proteins while DDA detected only four. We also targeted the same number of randomly selected human proteins and did not observe a single false-positive hit (Fig. S3). Thus, Picky enables detection of human proteins with high sensitivity and specificity.

SRM/PRM data is typically validated by monitoring the chromatographic coelution of multiple transitions for a given peptide[6]. This approach yielded convincing data for high amounts of spiked in proteins but somewhat unclear results for lower amounts (Fig. S4). We therefore

also compared the PRM data to annotated fragmentation spectra of corresponding synthetic peptides exported by Picky. The high similarity between the spectra (normalized spectrum contrast angle ≥ 0.5) further validated the PRM data (Fig. S5). We also compared all acquired UPS1-derived spectra with all fragmentation spectra in the Picky database (Figure S6). We did not observe a single false match with at least five transitions. Hence, Picky enables targeted protein identification with extremely high confidence.

In summary, the Picky tool (i) automatically generates optimized and scheduled SRM/PRM assays for proteins of interest and (ii) provides means to validate the data via known fragmentation spectra of corresponding synthetic peptides. Our benchmark experiment shows that Picky quickly generates an acquisition method that significantly outperforms non-targeted analysis. Picky thus greatly facilitates the targeted analysis of the human proteome.

1. Shi, T. *et al.* Advances in targeted proteomics and applications to biomedical research. *Proteomics* **16,** 2160–2182 (2016).
2. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol. Cell Proteomics* **11,** 1475–1488 (2012).
3. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Meth* **9,** 555–566 (2012).
4. Kusebauch, U. *et al.* Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **166,** 766–778 (2016).
5. Zolg, D. P., Wilhelm, M., Schnatbaum, K. & Zerweck, J. Building ProteomeTools based on a complete synthetic human proteome. *Nature* (2017). doi:10.1038/nmeth.4153
6. MacLean, B., Tomazela, D. M. & Shulman, N. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968 (2010).

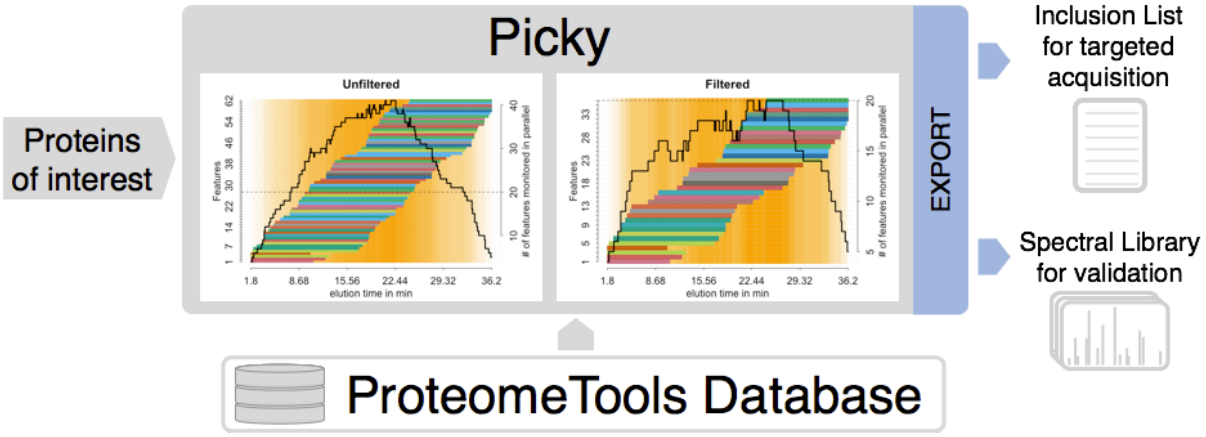

Fig. 1: Picky designs targeted acquisition methods (PRM/SRM) for proteins of interest by extracting data from pre-compiled ProteomeTools data. Filtering by the maximal number of co-eluting features selects the best set of peptides for the proteins of interest. Picky exports an inclusion list (for acquisition) and spectral information (for validation).

Supplemental Information

Fig. S1: Comparison between different available SRM or PRM method generators.

	Skyline	SRM atlas	Picky
SRM method generator	yes	yes	yes
PRM method generator	yes	no	yes
built-in library of synthetic spectra	no	yes	yes
scheduled acquisition	yes	yes	yes
user defined gradient	yes	no	yes
optimized scheduled acquisition	no	no	yes

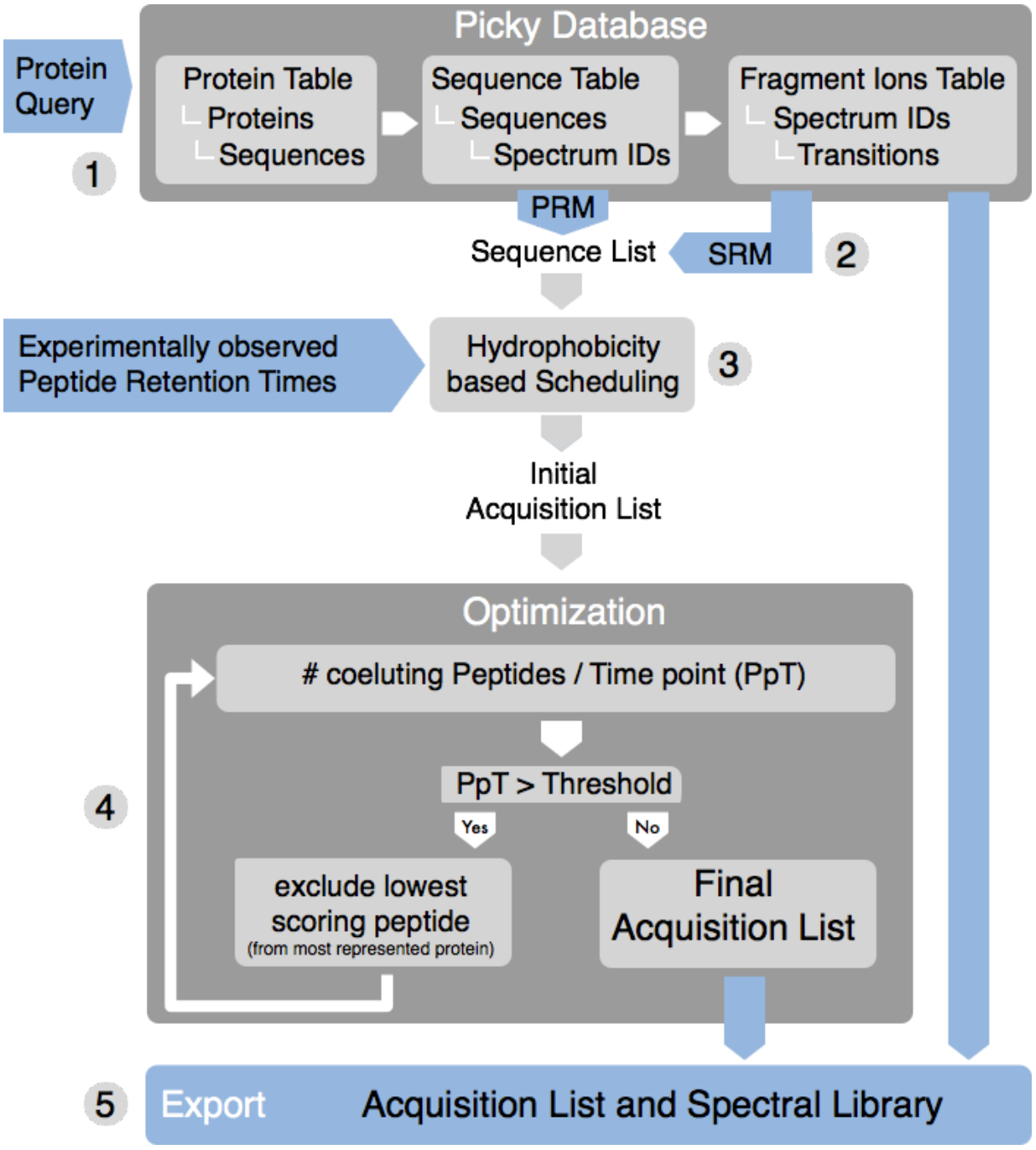

Fig. S2: Flowchart of the Picky Algorithm. For more details see section “Picky algorithm” in the Method description.

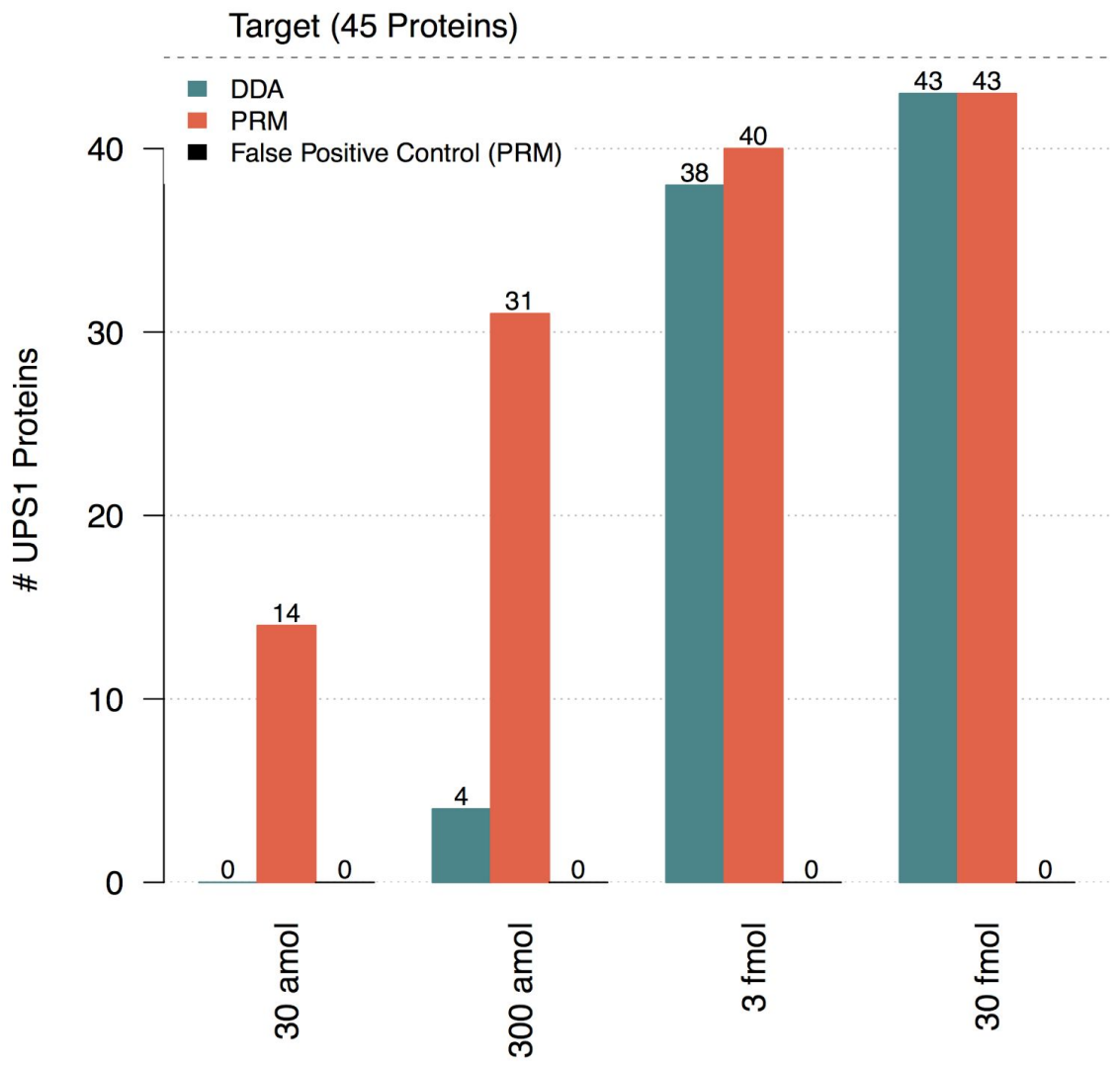

Fig. S3: Benchmark experiment to assess the specificity and sensitivity of PRM methods designed by Picky. As a reference sample different amounts of human proteins (UPS1) were spiked into 1.4 µg yeast lysate. A targeted method to detect all human proteins was designed by Picky (see Methods). To control false positives we targeted the same number of randomly selected human proteins (i.e. proteins not actually present in the sample). All samples were analyzed on the same Q Exactive Plus instrument via PRM and DDA. PRM markedly outperformed DDA without giving rise to false positive identifications. Note that we excluded three of the 48 human proteins in UPS1 since they share tryptic peptides with yeast proteins.

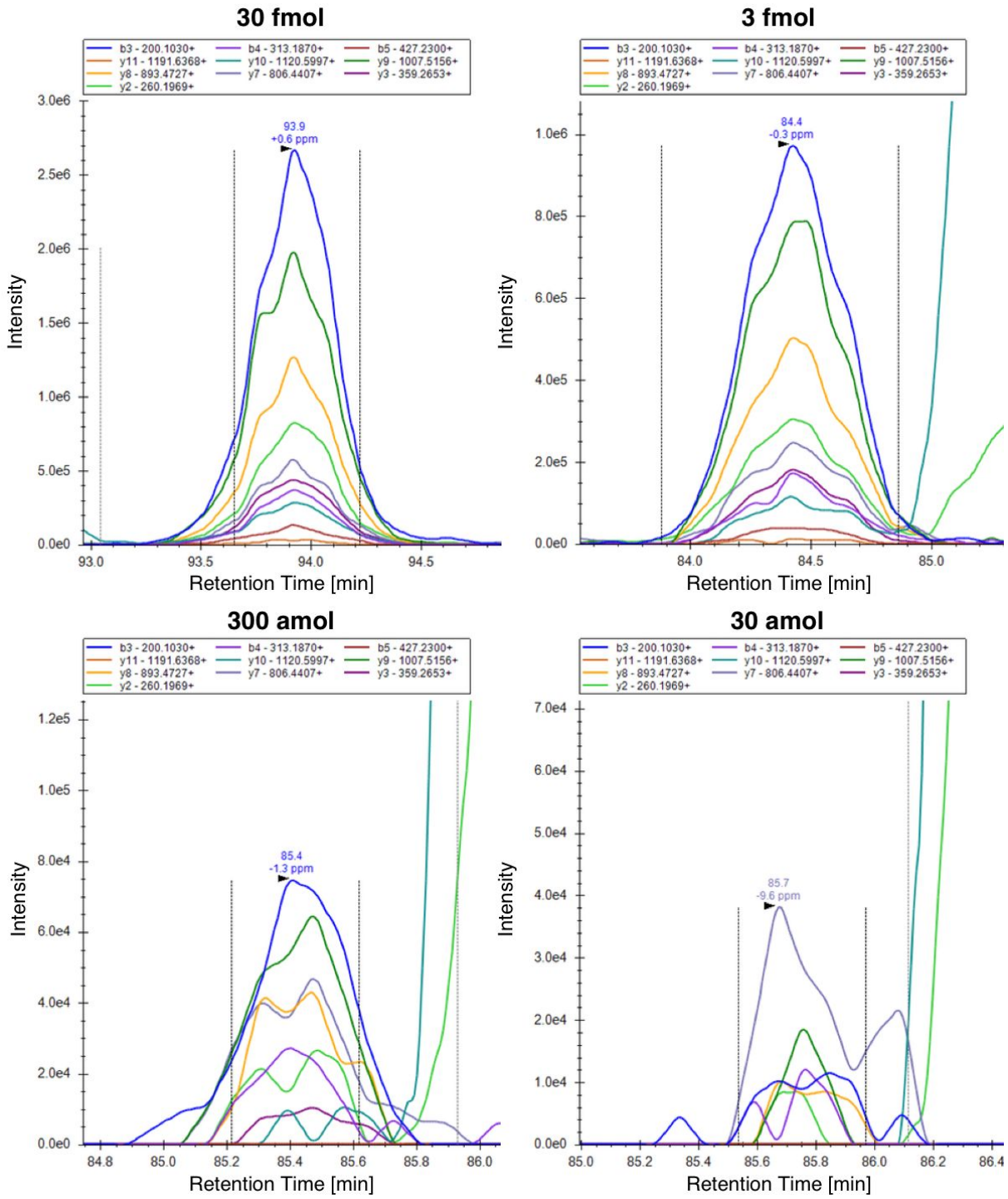

Fig. S4: Peaks of the peptide AGALNSNDAFVLK from the protein GSN. Figures were exported from Skyline for the four spike-in amounts 30 fmol, 3 fmol, 300 amol and 30 amol.

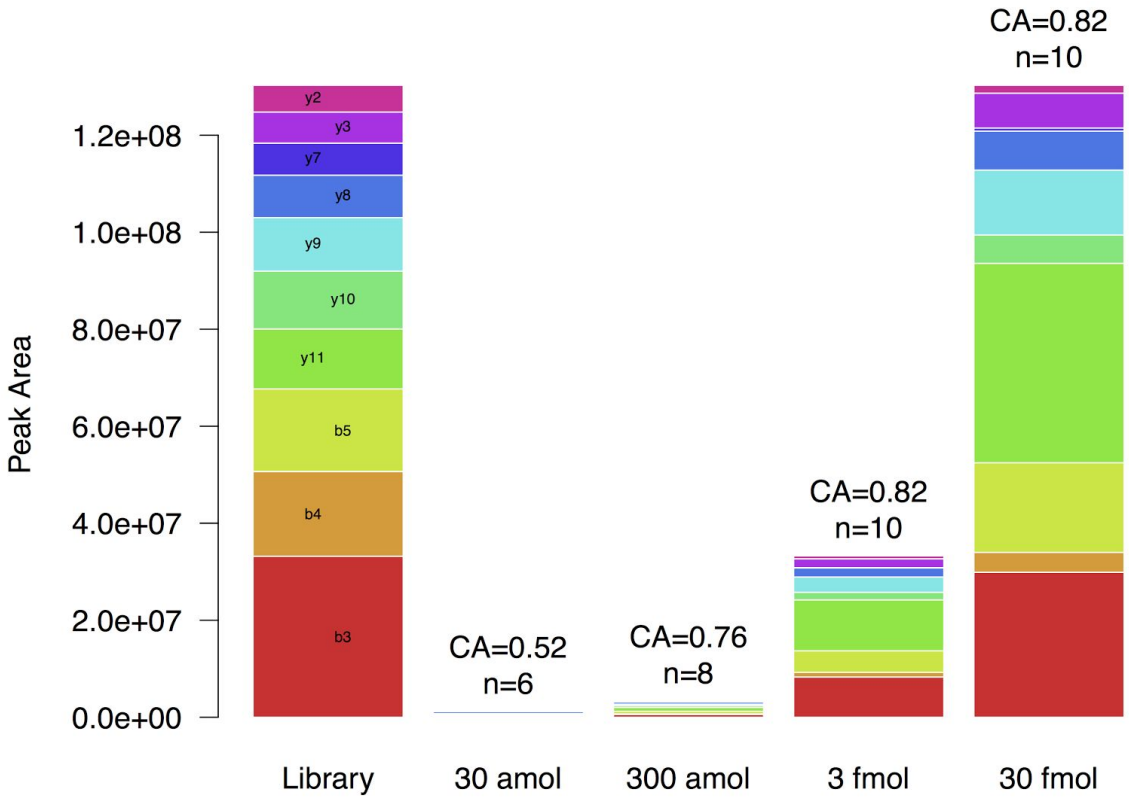

Fig. S5: Peak Areas of the peptide AGALNSNDAFVLK from the UPS1 protein GSN at different spike-in amounts (related to Fig. S4). The normalized spectrum contrast angle (CA) and the number of matched transitions is depicted above each stack and indicates spectrum similarity with the library spectrum. The different colors represent the different fragment ions. Library intensities were scaled to the maximal stack sum.

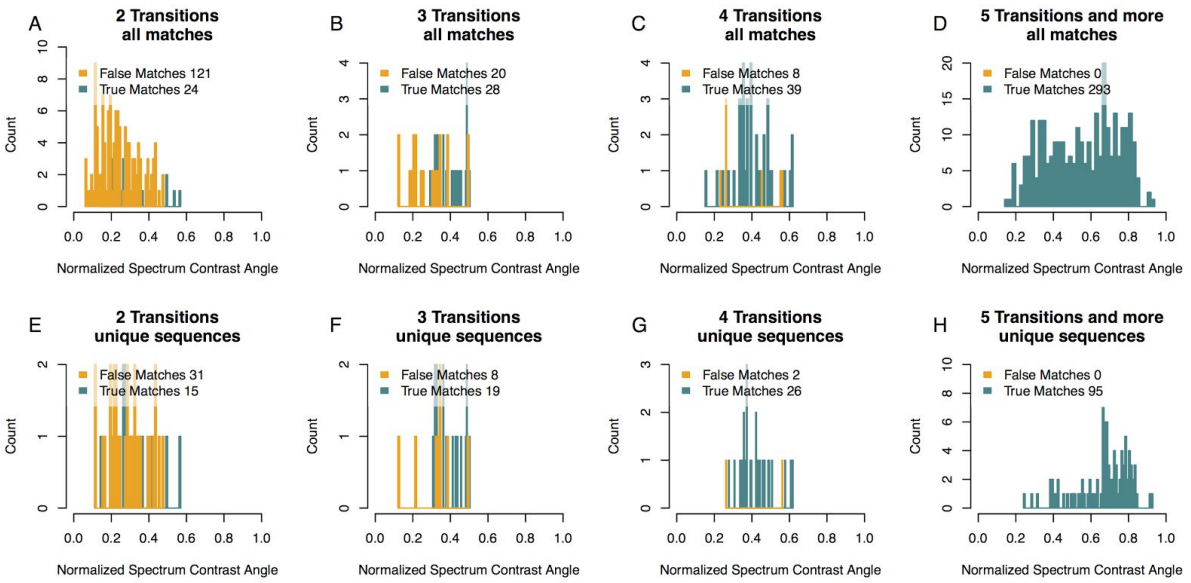

Fig. S6: Cross spectrum comparisons between the Picky Library and experimentally observed spectra from peptides of all UPS1 proteins in the benchmark dataset. The normalized spectrum contrast angle (CA) was calculated between spectra with matching precursor and transition masses (20 ppm mass accuracy). True and false matches for different numbers of transitions are shown (turquoise and orange, respectively). With at least five transitions no false match is observed. The top row shows results for all matches (A-D) while the bottom row depicts the highest CA for every unique sequence (E-H).

Method description

Picky Database

Data from ProteomeTools was precompiled using msms.txt text files from the available MaxQuant result files. For each peptide species and method-type the best scoring spectrum was picked. We found for almost all 19811 proteins listed in ProteomeTools at least one identification event in the provided msms.txt files while only 43 were without any identification event. Peptide species and method-types were distinguished by modification, charge, fragmentation type and collision energy. The corresponding raw fragmentation-spectra were extracted from raw-files with a python script using the Thermo MSFileReader and the MSFileReader.py bindings written by François Allen. The data was split into three tables holding information about proteins, peptides and corresponding transitions. All three tables were integrated into a SQLite database in R with the R-package RSQLite. The database is embedded in a shiny environment written in R to enable user friendly access. The R-code for Picky is available upon request.

Picky algorithm

Picky first collects all available peptide information for queried proteins considering the initial “Database Query” filters (fragmentation types, detector types, charge states, number of missed cleavages) and “Additional settings” filter (modified peptides, isoform specificity and proteotypic peptides; Fig S2-1). In case of SRM the highest intense transitions will be picked based on intensity and the set “Additional settings” filters (number of transitions and number of transitions with a m/z higher than the precursor m/z; Fig S2-2). Scheduling of the acquisition list is initialized by uploading a tab delimited table with a peptide-sequence and retention-time column (“Sequence” and “Retention Time”; Fig S2-3). This file can be obtained from any complex proteomic standard sample. Hydrophobicities of these sequences are calculated and fitted to the retention times using polynomial regression with the loess function as is implemented in R. Subsequently, peptides or transitions from queried proteins can be scheduled by predicting the retention time based on their hydrophobicity scores. The resulting “Initial Acquisition List” will be further optimized to fit the filter “Maximal number of in parallel monitored Features” in an iterative fashion (Fig S2-4). Among peptides that coelute and exceed the threshold of “Maximal number of in parallel monitored Features”, the lowest scoring peptide from the most represented protein(s) is removed from the acquisition list. It is important to note that Picky will remove proteins represented by only one remaining peptide in case no peptide from other proteins is scheduled at the corresponding retention time. Picky reports if proteins are excluded during the optimization procedure. To prevent this, users can either increase the maximal number of in parallel monitored features, decrease the retention time window (while increasing the risk of missing the peptide) or remove proteins from the query. The final acquisition list can be downloaded together with the corresponding spectra (Fig S2-4). The MaxQuant deconvoluted spectra and raw spectra are compiled into the MaxQuant msms.txt format. The MaxQuant deconvoluted msms.txt files can be as such imported into Skyline and used for Spectrum Comparison.

Sample Collection and Preparation

Universal Protein Standard 1 (UPS1) (Sigma Aldrich) was spiked at different amounts (30 amol, 300 amol, 3 fmol and 30 fmol) into 1.4 µg from total yeast protein extract. Yeast proteins were extracted from *S. cerevisiae* (strain BJ2168). Proteins were digested with trypsin and stage-tipped. Peptides were separated on a reverse phase HPLC system using a self packed column (ReproSil-Pur C18-AQ material; Dr. Maisch, GmbH; 3 h gradient; 5 to 75 % Acetonitrile). Peptides were ionized using an ESI source and analyzed on a Q-Exactive plus (Thermo Fisher). Samples were analyzed with a top10 data-dependent mode acquisition method (DDA) and parallel reaction monitoring method (PRM). Each UPS1 dilution was analyzed once for each analysis mode (DDA, PRM, PRM-False-Positive-Control) resulting in 12 samples. For DDA settings were briefly: Resolution 70 000 for MS1 (target value: 3,000,000 ions; maximum injection time of 20 ms); 17,500 for MS2 (maximum ion collection time of 60 ms with a target of reaching 1,000,000 ions; 2 Da isolation width). MS2 in PRM mode were acquired at a resolution of 17,500, AGC target at 200,000 ions, maximum injection time at 50 ms, isolation window 1.6 m/z). Inclusion lists with 118 transitions were obtained from Picky using default settings and querying all 48 UPS1 proteins. Maximal number of in parallel monitored features was set to 60 resulting in a cycle time between 3 and 4 seconds. A false positive control inclusion list was generated with Picky. 48 random human proteins different from the UPS1 set were queried in Picky and analyzed using the described settings. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[1] partner repository with the dataset identifier PXD007039.

Bioinformatic analyses

DDA runs were analyzed with MaxQuant 1.5.8.0[2] using default settings (multiplicity=0;Enzyme=Trypsin, including cut after proline; Oxidation (M) and N-terminal Acetylation set as variable modifications; carbamidomethylation (C) as fixed modification; database: uniprot yeast database from october 2014 and ups1 database as provided from Sigma Aldrich; Peptide and Protein FDR set to 0.01). UPS1 Proteins were defined as being identified if a proteinGroup listed a corresponding UPS1 protein at the first position. PRM data was analyzed with Skyline (3.6.0) with the following settings: Precursor Charges 2 to 7; ion charges 1 to 4; Ion types b and y; up to 6 product ions picked; auto-selection of matching transitions enabled; precursor m/z exclusion window = 2; ion match tolerance = 0.05 m/z; method match tolerance = 0.055 m/z; high selectivity extraction enabled; all matching scans were included; Resolving power of MS2 filtering was set to 17,500 at 400 m/z). A run specific spectral library was imported into Skyline using the peptide search import option. The msms.txt file was imported as downloaded from Picky. Each feature was manually validated in all samples by starting from the highest UPS1 spike in. Peaks needed to be in the range with the observed retention time in the highest concentrated UPS1 sample, have at least four matching transitions and a normalized spectral contrast angle (CA)[3] higher or equal to 0.5. All b and y ions as selected by Skyline were included into the calculation of the CA. Missing ions in recorded spectra were replaced with zero intensity. The observed median CA

was 0.8. Final results were exported as a transition report and compared with the proteinGroups.txt from the DDA analysis using the statistical computing language R. Proteins sharing selected peptides with *S. cerevisiae* or sharing a protein-group in the MaxQuant results were excluded from the analysis. Altogether, 45 UPS1 proteins were included in the final comparison.

1. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, **44(D1)**, D447–56 (2016).
2. Cox, J., & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26(12)**, 1367–1372 (2008).
3. Toprak, U. H. *et al.* Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Molecular & Cellular Proteomics*, **13(8)**, 2056–2071 (2014).