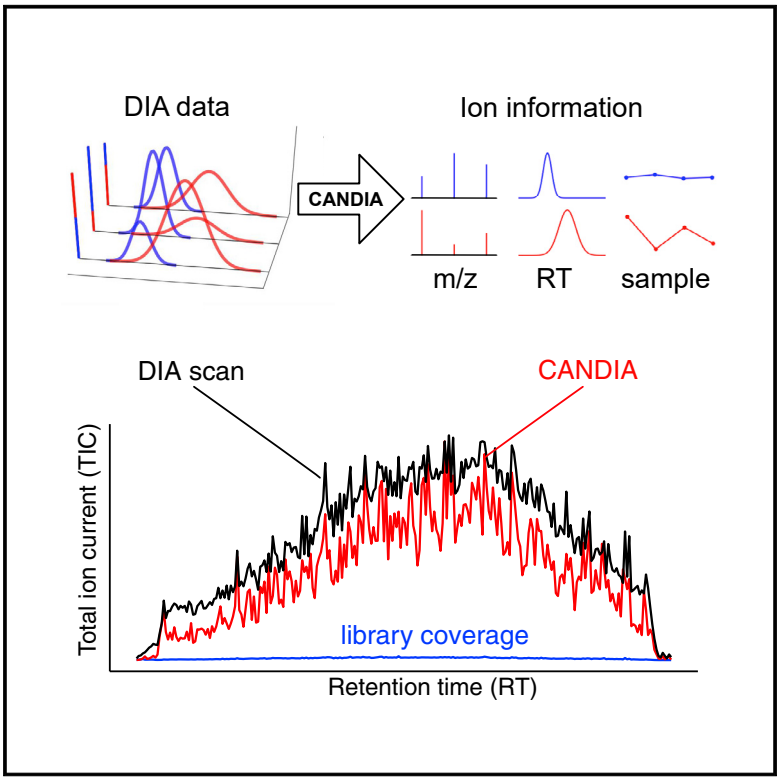
Parallel Factor Analysis Enables Quantification and Identification of Highly Convolved Data-Independent-Acquired Protein Spectra

Article

Patterns

Parallel Factor Analysis Enables Quantification and Identification of Highly Convolved Data-Independent-Acquired Protein Spectra

Graphical Abstract



Authors

Filip Buric, Jan Zrimec, Aleksej Zelezniak

Correspondence

aleksej.zelezniak@chalmers.se

In Brief

We developed a software pipeline that enables deep analysis of very large proteomics data in real time, complementing existing techniques with unbiased unsupervised tensor decomposition.

Highlights

- Conventional DIA spectral libraries cover less than 3% of a scan's total ion count
- CANDIA deconvolves peptide signals by leveraging all scan data
- CANDIA uses GPUs to enable tensor algebra on massive DIA mass spectrometry data
- CANDIA output enables high-confidence and precise quantitative proteomics

Buric et al., 2020, Patterns *1*, 100137
December 11, 2020
https://doi.org/10.1016/j.patter.2020.100137

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Parallel Factor Analysis Enables Quantification and Identification of Highly Convolved Data-Independent-Acquired Protein Spectra

Filip Buric,[1] Jan Zrimec,[1] and Aleksej Zelezniak[1,2,3,*]
[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg 412 96, Sweden
[2]Science for Life Laboratory, Tomtebodavägen 23a, Stockholm 171 65, Sweden
[3]Lead Contact
*Correspondence: aleksej.zelezniak@chalmers.se
https://doi.org/10.1016/j.patter.2020.100137

THE BIGGER PICTURE The latest high-throughput mass spectrometry-based technologies can record virtually all molecules from complex biological samples, providing a holistic picture of proteomes in cells and tissues and enabling an evaluation of the overall status of a person's health. However, current best practices are still only scratching the surface of the wealth of available information obtained from the massive proteome datasets, and efficient novel data-driven strategies are needed.

Powered by advances in GPU hardware and open-source machine-learning frameworks, we developed a data-driven approach, CANDIA, which disassembles highly complex proteomics data into the elementary molecular signatures of the proteins in biological samples. Our work provides a performant and adaptable solution that complements existing mass spectrometry techniques. As the central mathematical methods are generic, other scientific fields that are dealing with highly convolved datasets will benefit from this work.

1 **2** 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

High-throughput data-independent acquisition (DIA) is the method of choice for quantitative proteomics, combining the best practices of targeted and shotgun approaches. The resultant DIA spectra are, however, highly convolved and with no direct precursor-fragment correspondence, complicating biological sample analysis. Here, we present CANDIA (canonical decomposition of data-independent-acquired spectra), a GPU-powered unsupervised multiway factor analysis framework that deconvolves multispectral scans to individual analyte spectra, chromatographic profiles, and sample abundances, using parallel factor analysis. The deconvolved spectra can be annotated with traditional database search engines or used as high-quality input for *de novo* sequencing methods. We demonstrate that spectral libraries generated with CANDIA substantially reduce the false discovery rate underlying the validation of spectral quantification. CANDIA covers up to 33 times more total ion current than library-based approaches, which typically use less than 5% of total recorded ions, thus allowing quantification and identification of signals from unexplored DIA spectra.

INTRODUCTION

The ideal proteomic method should precisely quantify large sets of proteins across multiple samples. To this end, data-independent acquisition[1] (DIA) is an effective compromise between targeted proteomics using selected reaction monitoring (SRM) and label-free shotgun proteomics with data-dependent acquisition (DDA), combining the respective benefits of high accuracy and consistency[2,3] with high throughput.[4] Multiple issues are addressed, such as the inconsistent quantification due to stochasticity between runs, noticeable especially in DDA experiments with large sample series.[5,6] Despite this, an inherent issue is related to the exhaustive fragmentation of the specific mass range using defined isolation windows or "swaths."[7] Due to the width of these windows, fragment signals are highly overlapped or "convolved," with multiple precursors falling in the same

window, producing a set of highly overlapping ion mass spectra.[8–10] A computational solution to deconvolve such data would expand the coverage and efficacy of the DIA approach. Thus, development of novel data-analysis approaches is currently among major priorities in high-throughput proteomics.

The current standard approach for DIA analysis is targeted quantification of the acquired fragment data using spectral libraries containing fragmentation information for a particular peptide.[9–12] Library generation, however, is time consuming and specific to the instrument, chromatography, and experimental conditions, ideally requiring physical sample fractionation complemented with shotgun spectra acquisition.[13] Another limitation is that only a small portion of analytes are recovered, especially when library generation is based on data-dependent acquisition of relatively few selected high-intensity precursors.[14,15] Thus, the targeted search for DDA precursor fragments does not take full advantage of resulting digital records of all ions in scans generated in a data-independent manner.[7] Recent approaches based on large synthetic peptide libraries enable accurate prediction of peptide spectra directly from sequence data.[16,17] Computational approaches that utilize MS1-MS2 co-elution information to generate pseudo-spectra do not require the creation of experimental libraries.[18–20] These, however, suffer from the same overlapping fragment signal problem inherent to DIA, which is addressed using heuristics such as interference correction.[10,21,22]

Multiway tensor decomposition and other so-called matrix methods,[23,24] such as parallel factor analysis (PARAFAC), also called "canonical decomposition,"[25–28] use the entire acquired data to extract individual analyte signals and have been used for over four decades in mass spectrometry (MS) and other analytical technologies.[23,28–31] PARAFAC enables decomposition of multiway data arrays and facilitates the identification and quantification of independent underlying signals, termed "components," from convolved spectral data. Conveniently, DIA data can be naturally represented as a three-dimensional array or tensor, resulting from the linear combination of individual peptide mass spectra, their elution profiles, and their relative sample contribution, making it amenable to PARAFAC decomposition. However, given the sheer size of DIA proteomics datasets, where an experiment of 100 samples can easily generate more than half a terabyte of numerical data, computational decomposition of DIA proteomics data using conventional CPU-based multiway analysis frameworks become infeasible. Furthermore, existing PARAFAC applications usually involve smaller datasets consisting of at most a few hundred known analytes, so far limiting PARAFAC applications to relatively simple computational problems. On the other hand, with DIA proteomics data, one deals with an unknown set of tens of thousands of analytes, thus requiring a way to search a much larger model space than is currently achievable.

Hence, here we present a graphic processing unit (GPU)-accelerated multiway tensor decomposition approach called CANonical decomposition of Data-Independent Acquired spectra (CANDIA), consisting of a data decomposition pipeline that enables spectra retrieval and quantification of analytes directly from DIA data. By using a data-partitioning scheme and relying on the massive parallelism of modern GPU architectures, we achieve a technical leap, enabling untargeted decomposition of very large, high-throughput proteomics data. Moreover, the central method to the pipeline, PARAFAC, does not require *a priori* spectral information about the analytes in order to perform the decomposition. The individual deconvolved spectra produced by our pipeline may be analyzed with conventional peptide search engines[32–34] to produce peptide-spectrum matches (PSMs) for building high-accuracy spectral libraries. We show that, by using CANDIA, we can extract up to 33 times more analyte signal from DIA scans compared with library-based approaches, and, moreover, cover the entire *m/z* space of scans, enabling the usage of the majority of noise-accounted signal ions obtained from a sample. We also demonstrate that spectra recovered by CANDIA circumvent the problem of false quantifications, a major challenge present in targeted DIA proteomics.

RESULTS

CANDIA: A GPU-Accelerated Software Pipeline for Deconvolving DIA Data

We developed the CANDIA pipeline, capable of recovering spectral features in unsupervised fashion and computationally feasible time (Figure 1A), by leveraging the power of the modern tensor algebra frameworks PyTorch[35] and Tensorflow,[36] which take advantage of the parallelism and throughput of floating point operations in GPUs, as well as the distributed "big data" computing framework Apache Spark.[37] In brief, CANDIA partitions all provided DIA scans into a collection of small, independent tensors (including both MS1 and MS2 data), corresponding to precursor isolation windows and time intervals (Experimental Procedures P1). It then performs multiple decompositions of each of these tensors in parallel, accounting for a range of possible numbers of components in each tensor (Figure 1A and Experimental Procedures P2). As the observed intensities in DIA liquid chromatography-tandem MS scans result from linear combinations of individual fragmented peptide spectra, their elution profiles, and their relative abundance across all samples, each PARAFAC component ideally represents an analyte as a triplet of its *m/z* spectrum, retention time (RT) peak, and relative sample contribution (Figure 1B). The decomposition results are then refined by selecting the best models based on the quality of reconstructed signals, i.e., the unimodality of the elution profile (Experimental Procedures P3 and Note S1). A critical step in constructing a PARAFAC model is deciding *a priori* the number of components F, complicated by the fact that PARAFAC models do not "nest," i.e., a model for $F + 1$ is not simply a model for F with an extra component.[30] Deciding the value automatically is generally an open problem,[38] and the various diagnostics and procedures used to this end[39] often require human verification, which is not feasible for data-rich proteomics workflows, with hundreds of thousands of models that need to be examined. Our approach therefore exhaustively constructs all possible models within the configured range, then uses the shape of the resulting elution profiles and accounts for noise to automatically select valid models, resulting in optimal precursor identifications. The recovered *m/z* spectra (Figure 1C) can be directly searched using standard tool sets such as Crux,[40] TPP,[33] and MS-GF+[34] to (1) produce PSMs (Figure 1D), (2) build spectral libraries (Experimental Procedures P4), (3) be used for *de novo* sequencing, or (4) be used directly as linearly independent features for machine-learning applications.

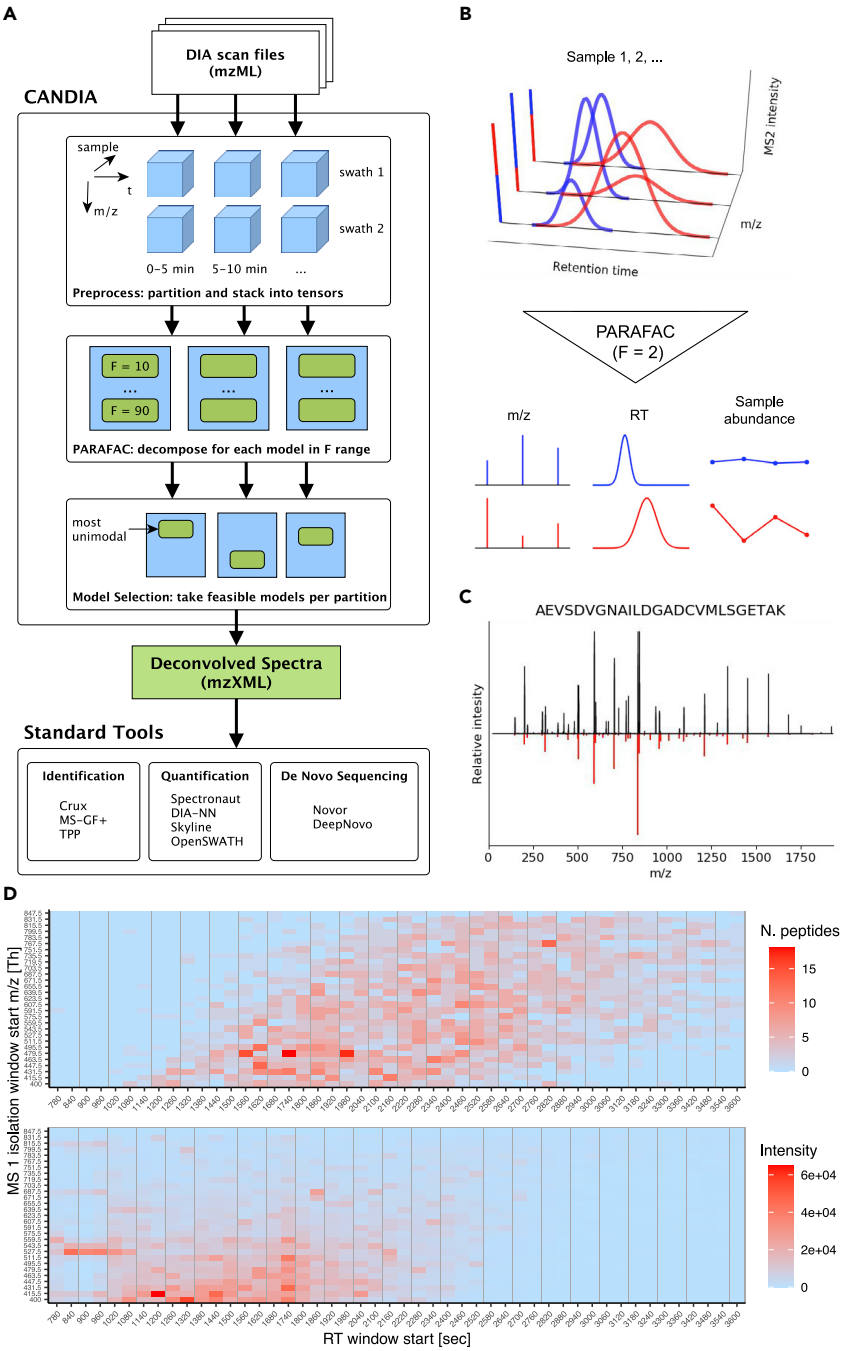


Figure 1. The CANDIA Pipeline, Illustration of PARAFAC Decomposition, and Example Results

(A) High-level structure of the CANDIA framework. Under the hood, CANDIA uses tools typically applied for processing of big data (on the order of hundreds of GB of numerical data) to perform PARAFAC decomposition of similarly large DIA data. It operates in a parallelized way, employing tensor computation frameworks that leverage the speed of GPU cards. CANDIA takes in all provided DIA scan files together, partitions them into a collection of independent tensors according to swath and retention time windows, then performs multiple decompositions of each of these tensors, accounting for a range of possible number of components, to account for an unknown number of peptides in each partition. The best models are selected such that most components have unimodal elution profiles (Experimental Procedures P3 and Figure 3). CANDIA output consists of a file in mzXML format containing the deconvolved spectra. This file is orders of magnitude smaller than the input scan files, which speeds up downstream analytical methods.

(B) Conceptual illustration of the PARAFAC decomposition method for two components. Acquired DIA MS1 and MS2 signals can be expressed as a linear combination of individual peptide mass spectra, their elution profiles, and their relative sample contribution. PARAFAC considers all sample scans at once and decomposes the three-dimensional (*m/z*, retention time [RT], sample) tensor structure into deconvolved components.

(C) Example of PARAFAC output spectrum matched to a peptide by Comet. Theoretical spectrum (predicted with Prosit)[16] of the peptide (black) is plotted against fragments matched (66%) to the deconvolved spectrum output by the pipeline (red).

(D) Peptide identification using Crux and MS-GF+ on CANDIA output (top) largely matches the distribution of input DIA scan MS1 intensities (bottom, single yeast lysate scan). Peptide count and intensities are shown per retention time (RT) and precursor isolation windows, according to the pipeline's data partitioning scheme. RT windows are highlighted by light-gray vertical lines. The horizontal streaks that show up in the ranges 11–17 min and 527–559 *m/z* (lower left) are likely contaminants (e.g., nothing was identified by Spectronaut [FDR <5%] in this range) and are not reflected in any CANDIA identifications.

Precise Protein Identification and Quantification with CANDIA

We first evaluated whether CANDIA-deconvolved spectra were identifiable by conventional peptide search engines. First, we tested our approach on a *Saccharomyces cerevisiae* lysate dataset,[2] referred to here as yeast replicates, which consisted of nine consecutive injections acquired in SWATH mode on a conventional Sciex 5600 QqTOF instrument with microflow setup. To identify the precursors, CANDIA solved a total of 176,175 PARAFAC models (29 swaths × 75 1-min windows × 81 models) on a GPU-equipped workstation. As a benchmark for comparison, we used DIA-Umpire,[18] a widely used tool for building spectral libraries directly from DIA data, and considered results from Crux (Comet coupled with Percolator) and MS-GF+ search engines separately, to assess their performance. While Crux identified a total of 2,014 proteins using the DIA-Umpire pseudo-spectra produced for each yeast replicate, in contrast only 684 proteins (1,553 peptides) were identified on the output from CANDIA. However, when considering only the proteins that appear in at least eight of the technical replicates, 583 proteins

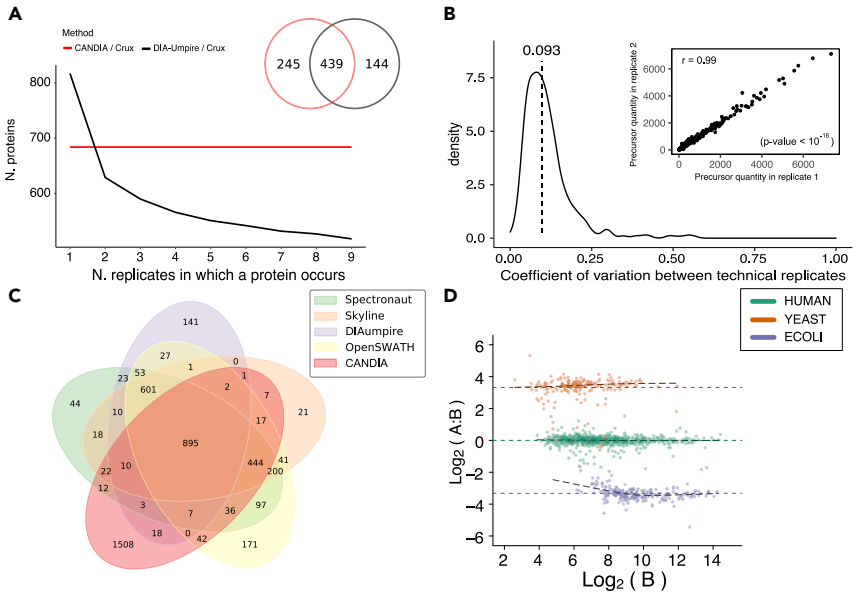


Figure 2. Precise Protein Identification and Quantification with CANDIA

(A) Proteins identified with Crux run on DIA-Umpire pseudo-spectra for each replicate, counted according to their prevalence across the replicates, compared with Crux results on CANDIA output. CANDIA produces deconvolved spectra from all input replicates, thus the number of IDs reflects the entire dataset. Inset: overlap of proteins identified by Crux with DIA-Umpire in at least eight replicates, and proteins identified with CANDIA.

(B) Precursor quantity coefficient of variation (median CV = 9.3%, plotted as dashed vertical line) across the yeast replicates dataset, obtained from DIA-NN using a CANDIA library. Inset: an example of highly correlated quantities between two replicates.

(C) Overlap between proteins identified with CANDIA coupled with Crux and MS-GF+ on the LFQbench HYE110 dataset, and published results from other tools. CANDIA results have 10-fold more unique identifications.

(D) LFQbench HYE110 results for CANDIA coupled with DIA-NN, showing quantification of human (green), *S. cerevisiae* (orange), and *E. coli* (purple) peptides. The DIA data are acquired from two hybrid proteome mixtures A and B with known organism concentrations. Plotted are log-transformed ratios ($\log_2$(A/B)) of peptide concentrations over the log-transformed intensity of sample B, against the expected values for each organism (horizontal dashed lines). Regression curves are shown as black dashed lines.

were found at 1% false discovery rate (FDR) using DIA-Umpire in conjunction with Crux (Figure 2A). The same pattern emerged for MS-GF+, only with far fewer identifications prevalent across the majority of replicates (see Table S1). PARAFAC decomposition, in contrast, captures the same analytes across all input samples, thus the commonality of high-confidence protein IDs across technical replicates is inherent to the method. Overall 53% overlap, consisting of 439 common proteins, was detected between CANDIA and DIA-Umpire coupled with Crux (considering only proteins identified in at least eight samples), with 245 proteins unique to CANDIA (Figure 2A). The peptide quantifications (Figure 2B) based on the reconstructed spectral library are precise (median coefficient of variation [CV] = 9.3% for the yeast replicates dataset) and reproducible (mean Pearson's r = 0.99 and $p < 1 \times 10^{-16}$ between the replicates).

The inconsistencies in peptide quantifications are often attributed to DDA approaches,[6] due to their stochastic nature of peptide selection, which is dependent on instrument performance. Although DIA methods typically produce far more complete data matrices,[5] the consistency of identification may be highly dependent on the inference correction procedures and the way FDR is estimated from DIA data. To exemplify this further, we built a library based on *in silico* digestion of a yeast proteome, whereby we randomly shuffled 30% of amino acids in each peptide sequence. Despite the fact that only peptides that did not exist in the original organism were considered, we quantified 2,691 and 642 proteins using the conventional library-based search tools Skyline[41] and DIA-NN,[10] respectively, at 1% FDR (Experimental Procedures P5). With the same search, Spectronaut[11] did not yield any identification. This is on average 185 times above the expected number of false discoveries (Figure S1). All of these identifications were attributed solely to the unique fragments arising from the mutated precursors, which would otherwise not be present in the original yeast spectral library and were extracted within predicted RT of the corresponding tool. Indeed, the score distributions of these sets of peptides suggest that the current decoy generation strategies do not provide a realistic model of the null hypothesis, which states that real peptide features are not different from the shuffled ones (Figure S2). In contrast, quantification with these tools using a library constructed from CANDIA output spectra only yielded identifications with one tool, and only 22 times higher than the expected number of false positives at 1% FDR. This indicates the validity of identifications when using PARAFAC-recovered spectra. To account for any confusion introduced to the inference algorithms of these software packages by providing them with completely spurious data, we also performed a sample-entrapment analysis[42] using a collection of *Archaea* proteomes as the entrapment partition. While generally robust to the false positives from the entrapment partition, the three software packages still yielded results from the *Archaea* proteomes, with the CANDIA library overall reducing false positives (Figure S3).

We next evaluated whether CANDIA could resolve spectra in a complex background such as the LFQbench HYE110 dataset,[43] containing two mixtures with different ratios of human, *Escherichia coli*, and *S. cerevisiae*, the latter two of which are present in quantities close to the limit of detection (5%) alternatively in the two mixtures. As, typically, the number of detected and quantified proteins is highly dependent on the particular tool and FDR estimation method, one would expect to find differences between available tools[43] and CANDIA, given that the latter works by searching against deconvolved spectra rather than by matching against scans using a library. Nevertheless, running Crux and MS-GF+ with an FDR threshold of 1% on CANDIA output spectra yielded a total of 3,024 proteins, comparable with the average 3,857 obtained by the other methods in the

benchmark study,[43] albeit at an overall lower rate of peptide identification (5,908 peptides with a median ratio of peptides to protein of 4) (Figure S4). Out of these proteins, 1,508 were unique to CANDIA (Figure 2C), which is 16 times higher than the median number of unique proteins of the other methods. We noticed that 601 proteins not identified with CANDIA were instead found using all methods that, in essence, use the same target-decoy fragment mass search algorithm for FDR estimation.[44] Analogously to the yeast example above, we built an *in silico* library of randomized peptide sequences, which resulted in a library with no peptides present in the non-randomized protein sequences of any of the three target organisms, as well as a sample-entrapment database using *Archaea* proteomes as the entrapment partition. For the randomized database, results were different between the tools, with up to 1,499 false protein identities found (Figure S5), 65 times higher than the expected number of false positives (23). In the sample-entrapment analysis, using the CANDIA library consistently reduced the number of entrapment proteins identified up to 100% (Figure S6). Given this underestimation of the false-positive rate made by the conventional tools, a considerable amount of the identifications at 1% FDR with tools other than CANDIA were thus put into the question. The quantification based on the CANDIA library, built using conventional methods[10,43] with an FDR threshold of 1%, resulted in precise ratios between the two mixtures A and B in the dataset (Figure 2D), showing precise quantification using a spectral library constructed directly from the decomposed spectra.

DIA Spectra Are a Molecular "Dark Matter"

Among the advantages of DIA is that acquired data represent a digital snapshot of all ions obtained from a sample.[7] While the idea is appealing, current library-based methods retrieve only a minor fraction of analytes present in a sample, leaving the majority of acquired DIA spectra unused. To demonstrate this, we calculated the overlap between *m/z* values of a DIA scan from the HYE110 dataset[43] and the corresponding published library. Allowing for a 5-min RT and a 50-ppm *m/z* tolerance, the spectral library matched 80.77% of all spectra in this scan. However, summing over the corresponding signal intensities, the library covered only 2.2% of total ion current (TIC) recorded in the scan (Figure 3A), showing that the majority of matched *m/z* points comprise baseline signals. Indeed, by filtering out all scan intensities below 1, the covered percentage of DIA spectra dropped to 3.46%. Thus, the remaining analyte signals, i.e., unlabeled features, are missed by current conventional tools that rely on targeted quantifications. On the contrary, the PARAFAC decomposition method considers virtually all *m/z* space (Experimental Procedures P2) in a DIA scan and discards unsystematic noise by leveraging the variability between scans to produce deconvolved spectra.[30] Thus, a reconstructed scan obtained by recombining PARAFAC output modes (Experimental Procedures P6) covered 72% of the same HYE110 scan TIC (Figure 3A), which is 33 times higher than what is covered by the spectral library and corresponds to more than twice (4.2%) the annotated spectral signal. Analogous results were obtained from *S. cerevisiae* lysate and its corresponding spectral library built using fractionation,[2] whereby the reconstructed PARAFAC pseudo-scan accounted for 12-fold more TIC than the corresponding library (Figure S7).

Extracting the realistic number of analytes, however, is dependent on the choice of the correct number of components.[30] The challenge is to identify the number of components that would best match the number of precursors in the samples. Finding the optimal number of components, however, is an NP-complete problem.[45] Thus, to minimize the risk of overfitting due to an overestimated number of components, we developed an effective empirical approach that functions under the assumption that every analyte has to elute as a single chromatographic peak, our so-called unimodality criterion (Experimental Procedures P3). Indeed, choosing the number of PARAFAC components with most unimodal elution profiles resulted in the optimal number of significant (FDR < 1%) precursor identifications and, importantly, this approach better captured the densest region of the scans (in terms of TIC), giving more confidence in the results (Figure 3B). We included all models with maximum unimodality (i.e., the entire feasible set) to account for the uncertainty in peptide-spectrum matching,[46] as it was observed that in denser scan regions (in terms of TIC) there are more models with maximum unimodality and, moreover, these recapitulate several of the same components (see Note S1). This is corroborated by theory, as models around the correct number of components will often not differ much in terms of loadings.[30] A large increase in number of components (with decreasing unimodality fraction) did not result in more peptide identifications; rather, it resulted in about half of the amount yielded by the optimal set. This is not surprising, as beyond the correct value, more components will start modeling noise or a single analyte may be captured by more non-physical components.[30] As an intuitive example, Figure 3C compares a pair of good and bad models for the yeast replicates dataset by showing the time modes (elution profiles) of each component in a single slice (single swath and time window), as output by PARAFAC. Examples of components from both these models are shown in Figure S8. As demonstrated, the good model captures proper elution curves, whereas non-optimal models comprise a significant number of fragmented or very narrow single-point curves, representing noise or split components that do not reflect any analytes. For the yeast lysate dataset, as the median of the unimodality fraction across the best models was 87% (Figure 3D), we attributed the remaining 13% of components to captured noise and elution curves clipped at the edge of the data slices, which were not counted by the peak detection routine (see Table S2).

As empirical evidence of the quality of CANDIA results, the total number of identified proteins for the yeast lysate dataset is comparable with library-based methods (Figures 2A and 2C), and the median number of recovered precursors per protein is eight, which is similar to what is typically expected from these approaches[43] (Figure S9). Accounting for imperfect deconvolutions, the majority (median across datasets = 85%) of recovered spectra (i.e., components) were not mapped to peptide space, and this remaining set of PARAFAC components is overall uncorrelated with the set of identified components (mean Pearson's $r = -7 \times 10^{-4}$) (Figure 3E). This was assessed by decomposing the dataset from Vowinckel et al.,[2] consisting of 30 yeast lysate samples from a study of the target of rapamycin (TOR) pathway, and calculating the correlation between identified and unidentified unimodal sample modes (the sample mode of a component holds the contribution of that component

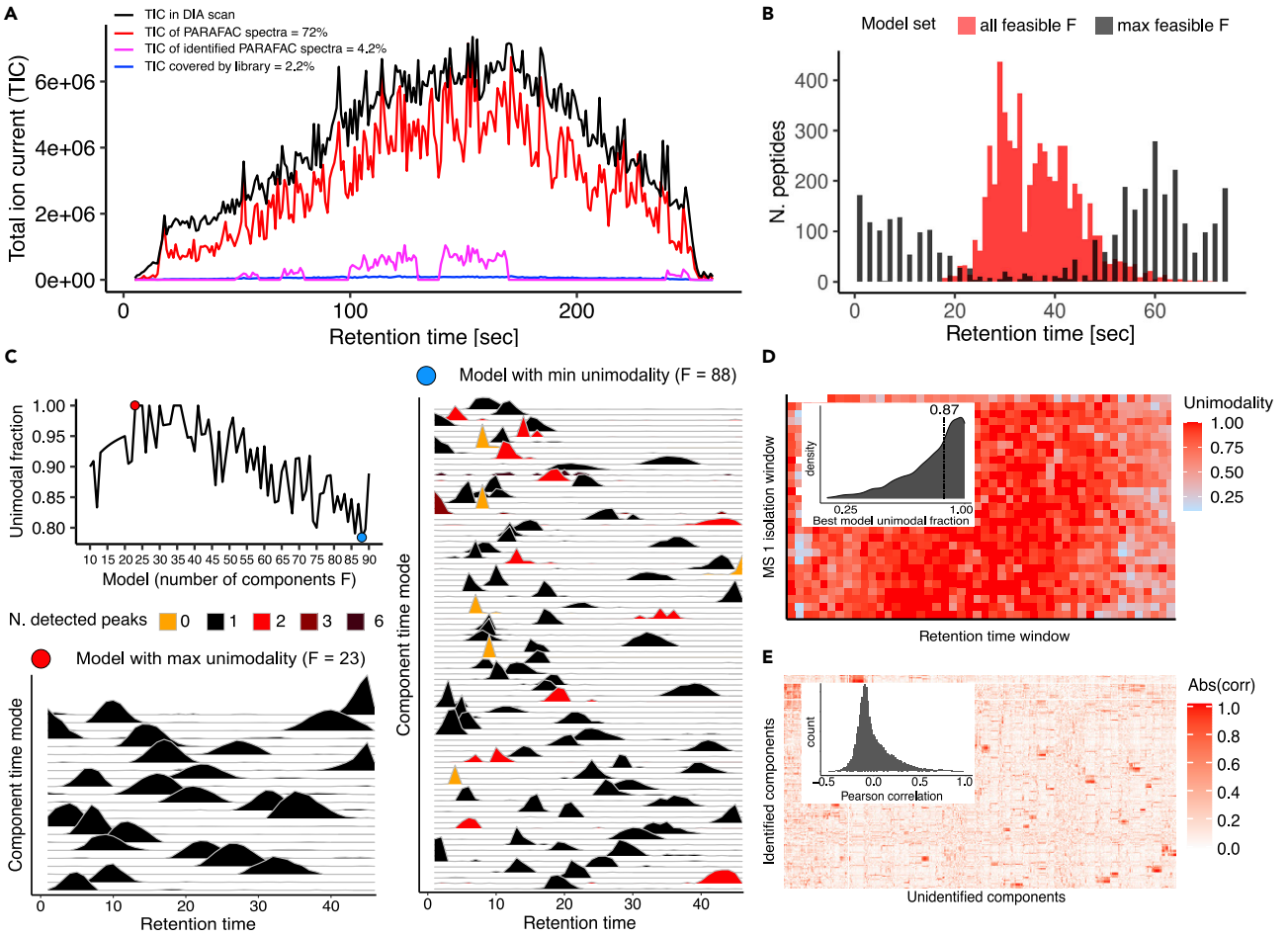


Figure 3. DIA Spectra Are Still Proteomic "Dark Matter"

(A) Typical DDA spectral library coverage of the total ion current (TIC) in a centroided DIA scan (HYE110), compared with that of recovered PARAFAC components and the subset of identified components. Retention time (RT) is binned by 30 s. Matching with the library allowed for a 5-min RT and a 50 ppm *m/z* tolerance, and at least four library fragments (product *m/z* points) needed to match for a spectrum to be considered covered. The scan reconstruction procedure is described in Experimental Procedures P6.

(B) Number of peptides identified by Crux (at 1% FDR) in the yeast replicates dataset (per 1-min time window), using the most unimodal models (1,553), compared with a control model set (1,111). The number of peptide matches is plotted when including all models with number of components chosen to maximize unimodality fraction (red), i.e., all feasible models. This is compared with matches against only the most complex feasible models (black). The distribution is similar for identifications using the least complex feasible models (Figure S11).

(C) Unimodality fractions of all models solved for an example slice of the yeast replicates dataset (MS1 isolation window 479–496 *m/z*, RT time window 29–30 min) are shown on top left. A good model (red dot) was chosen as $F = 23$, as it is completely unimodal. This model is compared with a poor model ($F = 88$, blue dot), with the lowest fraction (78%) of unimodal time components. One can see how some of the elution curves in the worst model are fragmented (higher count of detected peaks) or very narrow (even point-like, resulting in no peak detection). A slice spans 60 s (binned to a 1.3-s scan cycle) and the expected full-width at half maximum of a peak is 12 s.[2]

(D) Heatmap of maximum unimodality fraction across models for each slice, resulting from the decomposition of the yeast replicates dataset. Inset: distribution of best model unimodality fraction, with a median of 0.87.

(E) Heatmap of absolute Pearson correlations between sample modes of identified and unidentified components. Inset: the histogram of Pearson correlations.

to each sample, i.e., analyte abundance in sample), keeping only the most complex feasible model per slice. For this dataset, the average unimodal fraction was 90% and, by considering only these non-noisy components, we showed that unidentified components contain non-redundant information that could be leveraged by, for example, machine-learning approaches.[47,48]

Lastly, to extract information from the set of unidentified spectra using existing methods, we queried them for post-translational modifications (PTMs) using MS-GF+ and, moreover, performed *de novo* sequencing using the state-of-the-art machine-learning-based approach DeepNovo,[49] as well as the established tool Novor[50] (Experimental Procedures P7). We identified in total 186 PTMs in the non-enriched microflow yeast SWATH runs (Tables S3), which was twice more than using DIA-Umpire. Moreover, the latter exhibited very low prevalence of identifications, with only 8.2% appearing in at most two replicates (Figure S10). *De novo* sequencing benefits from deconvolved input.[51] Consequently, DeepNovo and Novor respectively

yielded 4 and 24 times more high-confidence (over 80% sequence correctness probability, see Experimental Procedures P7) *de novo* sequences from CANDIA output, compared with running on DIA-Umpire output (Note S2).

DISCUSSION

Here we presented CANDIA, a GPU-powered multiway decomposition framework enabling unsupervised and untargeted extraction of analyte signals from DIA data (Figure 1B). CANDIA solves thousands of decompositions in real time, enabling multiway analyses of dense data-independent-acquired spectra. Parallel factor analysis,[25,26] the multiway analysis technique behind CANDIA, takes advantage of cross-sample analyte variation, enabling deconvolution of mass spectra belonging to individual precursors (Figures 1B and 1C). The recovered spectra can then be searched using conventional peptide search engines[34,40] or *de novo* sequencing tools[49,50] to assign analyte identifications. Specifically, we demonstrated CANDIA quantification precision by building a library from recovered spectra and analyzing consecutive injections from yeast lysates acquired using a microflow setup.[2] We also showed CANDIA performance in quantifying complex background samples; that is, in an unsupervised fashion our framework identified peptides and their correct corresponding mixture quantity ratios with high confidence (peptide-level FDR <1%) in the LFQ benchmark HYE110 dataset[43] (Figure 2D). Apart from being a challenging benchmark from an acquisition point of view, as the samples contain a mixture of species in different ratios, correctly identifying peptides and mixture ratios is also not trivial from the data-analysis perspective: (1) it requires estimating the correct number of components, corresponding to the realistic number of analytes present in the sample; (2) the identification of high-quality recovered spectra is analogous to DDA, but directly from MS2 data without MS1 precursor mass isolation; (3) the quantification had to be correct for these recovered spectra in order to yield accurate ratios. Despite these challenges, our unsupervised framework yielded results similar to those of the established targeted methods (Figure 2).

The quantification and identification of specific analytes requires accurate estimation of FDRs, especially crucial when performing unsupervised and untargeted analyte quantification. For DIA data, the procedure is semi-targeted,[14] i.e., untargeted acquisition with the targeted analyte quantification either based on an experimental library or *in silico* methods.[52] Conversely, CANDIA does not depend on a library but instead builds one using recovered spectra from the observed data, with analyte identification performed post hoc using conventional peptide search engines. Comparison of CANDIA results with those of other methods showed substantial differences in protein identifications; for example, in the LFQbench HYE110 dataset 601 proteins were quantified by all other methods except CANDIA, whereas 1,508 proteins with at least one peptide were uniquely quantified by CANDIA using a library constructed from recovered DIA spectra. We considered that, despite the differences in FDR estimation procedures of benchmarked software, in essence they all use the similar target-decoy FDR estimation procedure.[44] This led to the hypothesis that the observed differences between CANDIA and other methods were due to the way FDR estimation is performed in targeted quantifications. Indeed, by randomly shuffling up to 30% of amino acids in peptide sequences and building *in silico* libraries for targeted approaches (such that none of the shuffled libraries shared precursor fragments with the experimental library, Experimental Procedures P5), and using these shuffled sequences as search databases to identify the recovered spectra from CANDIA, on average about 100-fold more false identities were reported by other tools when not using a CANDIA library (Figures S1 and S5). Moreover, by performing a sample-entrapment assessment using hybrid target organism-*Archaea* databases, a considerable amount of entrapment hits (false identities) were reported across tools and datasets (Figures S3 and S6). An explanation is that decoy generation techniques, such as random shuffling, sequence reversal, or introducing specific systematic mutations,[53,54] would generate overdiscriminating (overly sensitive) scores because of unrealistic fragmentation *m/z* values in decoys, compared with those present in natural proteomes (Figure S1). The resulting decoy score distributions calculated from DIA data thus allow even 30% mutated peptides to be identified as hits, as opposed to running search engines on deconvolved spectra output from CANDIA, which uses spectral properties instead of DIA data target-decoy features and results in more sensitive matching. Therefore, CANDIA can be used for building high-confidence spectral libraries directly from data, and these can also be used with other targeted approaches to prevent false identification. This singular example should of course be followed by further studies.

Furthermore, we found that CANDIA-recovered spectra contain on average twice as many confidently identified (1% FDR) post-translationally modified peptides than by using pseudo-spectra from established methods (Table S3). Although we consistently identified a total of only 101 modified peptides in our yeast lysate replicates dataset, these were identified directly from a regular chromatography setup without applying specialized PTM enrichment techniques.[55] We demonstrated that recovered spectra can also be *de novo* sequenced, using combinatorial and deep-learning approaches,[49,50] resulting in up to 24 times more peptide sequences.

Moving forward, we anticipate that further improvements to the framework will increase not only the quality and quantity of results but also running time. For this study, the method performed well due to the high-quality, robust chromatographic gradients in our datasets,[2,43] i.e., the yeast and HYE110 datasets had less than 5% average variability in RTs. To account for less-reproducible gradients, adding an RT alignment step[56,57] would certainly improve spectra recovery and, correspondingly, the number of peptide identifications, as this is crucial for PARAFAC to perform well, since the trilinearity assumption is no longer guaranteed to hold when shifts on the RT axis are present.[30] Additionally, a different decomposition method can potentially improve results, i.e., the theoretically better model in this case would be PARAFAC2, which allows for slight non-linearities in one mode (RT shifts in this case),[31,58] thus alleviating the requirement for robust gradients. However, at the time of this study no efficient and scalable implementation existed. At present the relative quantification is performed using a library constructed from deconvolved spectra, whereas the sample mode of each component already gives the relative contribution of that component

to each sample. In practice, we have seen this to be too imprecise to use for high-quality quantification. Thus, improvements to the decomposition would enable analyte quantification from sample modes directly. Our framework can be readily adapted to other types of DIA data, including sliding MS1 window techniques[4,59,60] and small-molecule metabolomics data.[61] As the pipeline relies on the high-level Python multiway framework TensorLy (compatible with major machine-learning backends),[62] it can be readily adapted to include additional separation dimensions, such as ion-mobility separation[63] using either four-way PARAFAC or Tucker3 decomposition. To conclude, as a state-of-the-art computational solution to deconvolve MS scans, CANDIA shows potential to greatly expand the coverage and efficacy of the DIA approach, and we hope it can serve as a general platform for multiway analysis of MS data.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Further information and requests for material and resources should be directed to and will be fulfilled by the Lead Contact, Aleksej Zelezniak (aleksej.zelezniak@chalmers.se).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

All source data used in this paper are from previously published studies and available on publically available repositories. We downloaded the following: (1) two yeast lysate datasets (9 technical replicates and 30 samples)[2,47] from ProteomeXchange: PXD010529; and (2) the HYE110 dataset[43] from ProteomeXchange: PXD002952.

The CANDIA pipeline is available on GitHub at https://github.com/fburic/candia (tag "submission").

Resource requirements and expected runtimes are described in P8. Pipeline Runtime.

P1. Preprocessing

DIA scans were partitioned and combined to form independent tensors for the decomposition stage. To determine the size of these partitions or "slices," we used the MS1 precursor isolation windows or "swaths" to cut the scans along the *m/z* axis, and a reasonable time window to cut the RT axis, depending on the chromatography. Partitioning according to swaths is a natural approach, as precursor-product spectra within one swath are independent of those in other swaths. For the yeast replicates and TOR study datasets the time window was chosen as 1 min, whereas for the HYE110 dataset, 5-min windows were taken. This choice balanced the number of expected elutants, due to differences in gradients (e.g., 20 min versus 40 min) and hence, the range of possible models, against the resulting number of slices. This RT partitioning is similar to the approach taken by iPLS,[64] although here it primarily serves to reduce model memory requirements and, secondarily, model complexity. We saw that results are fairly robust to different window sizes (see Table S2), although too-narrow windows will increase the number of clipped (partial) elution peaks while too-large windows increase the complexity of the models, leading to a slight drop in quality of the final results. Future improvement of the approach ought to include a less arbitrary choice.

To facilitate processing, we converted the scan files to tabular format and partitioned them in parallel using the pyteomics package[65] and distributed computing framework Apache Spark. Each such slice thus contained the same (*m/z*, RT) partition for all input scan files, which were then "stacked," resulting in a (*m/z*, RT, sample) tensor structure, encoded as a NumPy[66] array. Technically, each MS1 survey scan and its respective MS2 spectra were aligned along the time axis, as they ought to form a single variable, i.e., the same column in the resulting matrix. The preprocessing step resulted in a collection of independent tensors that span the entire *m/z* and RT range of each sample. For more details, see Note S3.

P2. PARAFAC Decomposition

Each slice tensor ($\underline{\mathbf{D}}$) resulting from the preprocessing step was decomposed using PARAFAC into a sample mode *S*, a (retention) time mode *T*, and an *m/z* mode *M*, plus a residual error term $\underline{\mathbf{E}}$, for a given number of components *F* spanning a predetermined range. For an explicit form using the Kruskal operator,[28] see Equation 1:

$$\underline{\mathbf{D}} = [[S, T, M|F]] + \underline{\mathbf{E}}, \text{for } F = 10, ..., 90. \quad \text{(Equation 1)}$$

Each of these three mode matrices consist of *F* components, which correspond to separable analyte signals. That is, assuming perfect decomposition, each column in *S*, each column in *M*, and each row in *T* corresponds to the *m/z* spectrum, elution profile, and sample contribution, respectively of a single peptide. As this number *F* is unknown *a priori*, we performed the decomposition for an expected number of peptides within a slice. The choice of *F* value range was informed by inspecting a scan with Spectronaut. A non-negativity constraint was imposed on all modes, a natural assumption as they model non-negative physical quantities (concentrations and particle counts). Additionally, the search for a solution is more efficient[30] and prevents obtaining negative profiles due to imperfect decomposition (from noise or low variability).[27,67] All slice tensors were decomposed in parallel using the TensorLy GPU-adapted implementation.,[62] with PyTorch as a backend.

P3. Model Selection

To select the best model per slice from the range generated in the previous step, we counted the peaks of the time mode of each component of each model using a continuous wavelet transformation approach,[68] implemented in the SciPy package.[69] As each analyte should have a single elution peak, we counted, per model, the fraction of components with a single peak. Among all models generated for a slice, we chose all models with maximum fraction of unimodal time modes (see Figure 3C for an illustration). To test the performance of this criterion, we constructed spectra files from two other model sets by choosing only the smallest and largest maximally unimodal *F*, respectively, for each slice. These spectra files were then analyzed with Crux (Comet and Percolator) and the number of high-confidence peptide identifications was compared. Including all models with highest unimodality performed better (Figure 3B).

P4. Identification and Quantification

The spectra from the best models are saved to an mzXML file. This resembles a DDA file, since each scan entry consists of the MS2 part of the deconvolved spectrum along with its corresponding highest-intensity MS1 peak as precursor (all MS1 points are included, however). This file was then searched using Comet and MS-GF+ to produce PSMs in conjunction with a proteome FASTA database, using the same mass tolerance as the initial acquisition (i.e., 40 ppm for the yeast replicates and TOR study datasets, and 50 ppm for the HYE110 dataset). Comet results were then filtered using Percolator at 1% FDR. The confidence assessment for both MS-GF+ and Percolator was done using reversed decoys. To search for PTMs, we preconfigured MS-GF+ to account for acetylations, succinylations, phosphorylations, and core 1 GalNAc glycosylations as variable modifications anywhere in the peptide, allowing for maximum 384 variable modifications.

The output CANDIA mzXML file was used to construct a spectral library following the protocol in Schubert et al.,[13] using Comet as a source of PSMs. DIA-NN was then run on the initial DIA scan files using this library to produce peptide quantities, using default parameters. Benchmark results for the HYE110 dataset were obtained using the lfqbench R package.[43]

P5. False-Positive Assessment

Proteome FASTA database files were randomized such that 30% of each trypsin-digested protein sequence was shuffled. DIA-NN was used as the source of the *in silico* library constructed from the shuffled databases for all three software packages. It uses a spectral library built using fractionation[2] as training input to predict RT values. Baseline results were obtained using published spectral libraries.[2,43] For the sample-entrapment trial, a set of *Archaea* proteomes (see Note S4) were concatenated to the original organism databases. The size of these *Archaea* entrapment partitions were of approximately equal size to the sample (organism) partitions.

To assess the effect of using CANDIA as a preprocessor, we ran all tools using a spectral library created from the CANDIA-deconvolved spectra and the corresponding shuffled database, following the protocol in Schubert et al.[13] For all these runs we removed any resulting spectra that were also found in the published libraries, thus ensuring we generated exclusively false positives. In terms of parameters, Skyline was run analogously with Navarro et al.,[43] Spectronaut, and DIA-NN with default settings, except with a 100% FDR threshold to allow selecting results at different false discovery levels.

P6. Scan Reconstruction from a PARAFAC Model

The output PARAFAC sample mode *S*, RT mode *T*, and *m/z* mode *M*, consisting of *F* components resulting from the decomposition of a dataset, were used to reconstruct a pseudo-scan comprising the deconvolved analytes. A pseudo-scan P_i is an (*m/z*, RT) matrix corresponding to the input DIA scan *i* in the dataset. It is obtained by summing over the outer products of the *m/z* mode *m* and RT mode *t*, multiplied by contribution to scan *i* from the sample mode *s*, for all unimodal PARAFAC components *r* in the model.

$$P_i = \sum_{r=1}^{F} (m_r \otimes t_r) \cdot s_r(i) \text{for input sample scan } i. \quad \text{(Equation 2)}$$

Lastly, the resulting intensities in P_i were scaled back to the values in the corresponding scan *i*, and multiplied by the model coefficient of determination R^2,as PARAFAC solutions do not preserve the scaling of the input tensor.[30] This is, in effect, the reverse operation to PARAFAC for a single input sample, discarding the residuals (Equation 1 and Note S1; Equation 2). The above operation was done piecewise for each independent tensor produced by CANDIA partitioning of the input dataset.

P7. *De Novo* Sequencing

The output CANDIA mzXML was converted to MGF format and subsequently set as input to Novor[50] and DeepNovo.[49] Novor was run using a mass tolerance of 50 ppm and DeepNovo a tolerance of 10 ppm. Novor was set to collision-induced dissociation fragmentation and time-of-flight mass analyzer, using otherwise default parameters. For DeepNovo, the pretrained *yeast.low.coon_2013* model was used, with a beam size of 5. For the baseline DIA-Umpire results, only the highest-quality (Q1) extracted features were used, since these are far more likely to lead to good sequencing results, as good fragment coverage is needed.[51] Moreover, we considered only sequences that appear in at least six out of nine replicates based on DIA-Umpire features. Both tools assign a probability of correctness to the sequences, and 80% would be an acceptable threshold.[49,50]

P8. Pipeline Runtime

CANDIA itself requires an estimated 7 h (out of which the decomposition step takes an estimated 6 h) and downstream tools an additional 1 h, assuming availability of computing resources (see Table S4 for a breakdown of pipeline steps). The current work was performed on a workstation with 20 CPUs at 3.3 GHz and 64 GB RAM; however, the preprocessing and downstream steps perform well on as few as 8 CPUs and 16 GB RAM. For the decomposition step, two NVIDIA V100 GPU cards with 32 GB of RAM were used with highest performance. Alternatively, two workstation GP100 cards with 16 GB of RAM were used during development, with about half the performance. Given the granularity of the processing, as little as 8 GB of GPU RAM would still perform acceptably.

There is much that can be sped up starting from the current prototype, especially for the decomposition, by reducing the range of models (the primary factor in computation scaling) and optimizing the decomposition code. The other methods presented here took less time to compute (1–4 h, depending on the task), although we reiterate that they only examine a subset of the input data, whereas PARAFAC processes it in its entirety. Note, however, that in certain use cases, the simplified CANDIA output enables much faster analyses. For example, including multiple PTMs in an MS-GF+ search, the total runtime with CANDIA was about 24 h whereas with DIA-Umpire it was about 72 h (out of which the MS-GF+ search took 70 h), and more hits were obtained with CANDIA (Table S3).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100137.

ACKNOWLEDGMENTS

The computations were mainly performed on resources at the Chalmers Center for Computational Science and Engineering (C3SE), and partially at the High Performance Computing Center North, provided by the Swedish National Infrastructure for Computing. Mikael Öhman at C3SE is acknowledged for assistance concerning technical aspects in making the code run on the C3SE resources. We thank Hieu Tran for assistance in running DeepNovo, Vadim Demichev for assistance in running DIA-NN, and Lukas Reiter (Biognosys) for providing access to a trial version of the Spectronaut software. We thank Tejas Gandhi for explaining details of Spectronaut software usage. We also thank Kate Campbell for useful discussions and early feedback on the manuscript. F.B., J.Z., and A.Z. were supported by SciLifeLab funding.

AUTHOR CONTRIBUTIONS

Conceptualization, F.B. and A.Z.; Methodology, F.B. and A.Z.; Software, F.B. and J.Z.; Validation, F.B. and A.Z; Formal Analysis, F.B.; Investigation, F.B.; Resources, F.B., J.Z., and A.Z.; Data Curation, F.B. and J.Z.; Writing – Original Draft, F.B. and A.Z.; Writing – Review & Editing, F.B., J.Z., and A.Z.; Visualization, F.B.; Supervision, A.Z.; Project Administration, F.B. and A.Z.; Funding Acquisition, A.Z.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 25, 2020
Revised: September 14, 2020
Accepted: October 12, 2020
Published: November 5, 2020

REFERENCES

1. Venable, J.D., Dong, M.-Q., Wohlschlegel, J., Dillin, A., and Yates, J.R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods *1*, 39–45.
2. Vowinckel, J., Zelezniak, A., Bruderer, R., Mülleder, M., Reiter, L., and Ralser, M. (2018). Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. Sci. Rep. *8*, 4346.
3. Rosenberger, G., Liu, Y., Röst, H.L., Ludwig, C., Buil, A., Bensimon, A., Soste, M., Spector, T.D., Dermitzakis, E.T., Collins, B.C., et al. (2017). Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. Nat. Biotechnol. *35*, 781–788.
4. Messner, C., Demichev, V., Bloomfield, N., Ivosev, G., Wasim, F., Zelezniak, A., Lilley, K., Tate, S., and Ralser, M. (2019). ScanningSWATH enables ultra-fast proteomics using high-flow chromatography and minute-scale gradients. bioRxiv. https://doi.org/10.1101/656793.
5. Collins, B.C., Hunter, C.L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S.L., Chan, D.W., Gibson, B.W., Gingras, A.-C., Held, J.M., et al. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nat. Commun. *8*, 291.
6. Zhang, B., Käll, L., and Zubarev, R.A. (2016). DeMix-Q: quantification-centered data processing workflow. Mol. Cell. Proteomics *15*, 1467–1478.
7. Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). 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*, https://doi.org/10.1074/mcp.O111.016717.
8. Pappireddi, N., Martin, L., and Wühr, M. (2019). A review on quantitative multiplexed proteomics. Chembiochem *20*, 1210–1224.
9. Peckner, R., Myers, S.A., Jacome, A.S.V., Egertson, J.D., Abelin, J.G., MacCoss, M.J., Carr, S.A., and Jaffe, J.D. (2018). Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. Nat. Methods *15*, 371–378.

10. Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S., and Ralser, M. (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat. Methods *17*, 41–44.

11. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinović, S.M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., et al. (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol. Cell Proteomics *14*, 1400–1410.

12. Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. *32*, 219–223.

13. Schubert, O.T., Gillet, L.C., Collins, B.C., Navarro, P., Rosenberger, G., Wolski, W.E., Lam, H., Amodei, D., Mallick, P., MacLean, B., et al. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat. Protoc. *10*, 426–441.

14. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B.C., and Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol. Syst. Biol. *14*, https://doi.org/10.15252/msb.20178126.

15. Deutsch, E.W., Perez-Riverol, Y., Chalkley, R.J., Wilhelm, M., Tate, S., Sachsenberg, T., Walzer, M., Käll, L., Delanghe, B., Böcker, S., et al. (2018). Expanding the use of spectral libraries in proteomics. J. Proteome Res. *17*, 4051–4060.

16. Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., et al. (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat. Methods *16*, 509–518.

17. Gabriels, R., Martens, L., and Degroeve, S. (2019). 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. *47*, W295–W299.

18. Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C., and Nesvizhskii, A.I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods *12*, 258–264, 7 p following 264.

19. Wang, J., Tucholska, M., Knight, J.D.R., Lambert, J.-P., Tate, S., Larsen, B., Gingras, A.-C., and Bandeira, N. (2015). MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. Nat. Methods *12*, 1106–1108.

20. Li, Y., Zhong, C.-Q., Xu, X., Cai, S., Wu, X., Zhang, Y., Chen, J., Shi, J., Lin, S., and Han, J. (2015). Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. Nat. Methods *12*, 1105–1106.

21. Bao, Y., Waldemarson, S., Zhang, G., Wahlander, A., Ueberheide, B., Myung, S., Reed, B., Molloy, K., Padovan, J.C., Eriksson, J., et al. (2013). Detection and correction of interference in SRM analysis. Methods *61*, 299–303.

22. Keller, A., Bader, S.L., Shteynberg, D., Hood, L., and Moritz, R.L. (2015). Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition mass spectrometry (MS) using SWATHProphet. Mol. Cell Proteomics *14*, 1411–1418.

23. Likić, V.A. (2009). Extraction of pure components from overlapped signals in gas chromatography-mass spectrometry (GC-MS). BioData Min *2*, 6.

24. Bevilacqua, M., Bro, R., Marini, F., Rinnan, Å., Rasmussen, M.A., and Skov, T. (2017). Recent chemometrics advances for foodomics. Trends Analyt. Chem. *96*, 42–51.

25. Harshman, R.A. (1970). Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multimodal Factor Analysis (University of California at Los Angeles).

26. Carroll, J.D., Douglas Carroll, J., and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika *35*, 283–319.

27. Bro, R. (1997). PARAFAC. Tutorial and applications. Chemometrics Intellig. Lab. Syst. *38*, 149–172.

28. Kolda, T.G., and Bader, B.W. (2009). Tensor decompositions and applications. SIAM Rev. *51*, 455–500.

29. Gorrochategui, E., Jaumot, J., Lacorte, S., and Tauler, R. (2016). Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow. Trends Analyt. Chem. *82*, 425–442.

30. Smilde, A., Bro, R., and Geladi, P. (2005). Multi-way Analysis: Applications in the Chemical Sciences (John Wiley & Sons).

31. Johnsen, L.G., Skou, P.B., Khakimov, B., and Bro, R. (2017). Gas chromatography-mass spectrometry data processing made easy. J. Chromatogr. A *1503*, 57–64.

32. Park, C.Y., Klammer, A.A., Käll, L., MacCoss, M.J., and Noble, W.S. (2008). Rapid and accurate peptide identification from tandem mass spectra. J. Proteome Res. *7*, 3022–3027.

33. Deutsch, E.W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., and Moritz, R.L. (2015). Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin. Appl. *9*, 745–754.

34. Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. *5*, 5277.

35. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic Differentiation in PyTorch (openreview.net). https://openreview.net/pdf/25b8eee6c373d48b84e5e9c6e10e7cbbbce4ac73.pdf.

36. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv, 1603.04467 [cs.DC].

37. Zaharia, M., Xin, R.S., Wendell, P., and Das, T. (2016). Apache spark: a unified engine for big data processing. Commun. ACM *59*, https://doi.org/10.1145/2934664.

38. Liu, K., da Costa, J.P.C.L., So, H.C., Huang, L., and Ye, J. (2016). Detection of number of components in CANDECOMP/PARAFAC models via minimum description length. Digit. Signal. Process. *51*, 110–123.

39. Bro, R., and Kiers, H.A.L. (2003). A new efficient method for determining the number of components in PARAFAC models. J. Chemom. *17*, 274–286.

40. McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C.E., Diament, B., Frewen, B., Howbert, J.J., Hoopmann, M.R., Käll, L., Eng, J.K., et al. (2014). Crux: rapid open source protein tandem mass spectrometry analysis. J. Proteome Res. *13*, 4488–4491.

41. Pino, L.K., Searle, B.C., Bollinger, J.G., Nunn, B., MacLean, B., and MacCoss, M.J. (2017). The Skyline ecosystem: informatics for quantitative mass spectrometry proteomics. Mass Spectrom. Rev. *39*, 229–244.

42. Granholm, V., Noble, W.S., and Käll, L. (2011). On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. J. Proteome Res. *10*, 2671–2678.

43. Navarro, P., Kuharev, J., Gillet, L.C., Bernhardt, O.M., MacLean, B., Röst, H.L., Tate, S.A., Tsou, C.-C., Reiter, L., Distler, U., et al. (2016). A multicenter study benchmarks software tools for label-free proteome quantification. Nat. Biotechnol. *34*, 1130–1136.

44. Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Brusniak, M.-Y., Hengartner, M.O., and Aebersold, R. (2011). mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat. Methods *8*, 430–435.

45. Håstad, J. (1989). Tensor rank is NP-complete. In Automata, Languages and Programming, G. Ausiello, M. Dezani-Ciancaglini, and S. Ronchi Della Rocca, eds. (Springer), pp. 451–460.

46. Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J. Proteome Res. *7*, 29–34.

47. Zelezniak, A., Vowinckel, J., Capuano, F., Messner, C.B., Demichev, V., Polowsky, N., Mülleder, M., Kamrad, S., Klaus, B., Keller, M.A., et al.

(2018). Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knockouts. Cell Syst *7*, 269–283.e6.

48. Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., and Ralser, M. (2017). Designing and interpreting "multi-omic" experiments that may change our understanding of biology. Curr. Opin. Syst. Biol. *6*, 37–45.
49. Tran, N.H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017). De novo peptide sequencing by deep learning. Proc. Natl. Acad. Sci. U. S. A. *114*, 8247–8252.
50. Ma, B. (2015). Novor: real-time peptide de novo sequencing software. J. Am. Soc. Mass Spectrom. *26*, 1885–1894.
51. Muth, T., and Renard, B.Y. (2018). Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief. Bioinform *19*, 954–970.
52. Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., and Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat. Commun. *11*, 146.
53. Wang, G., Wu, W.W., Zhang, Z., Masilamani, S., and Shen, R.-F. (2009). Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. Anal. Chem. *81*, 146–159.
54. Levitsky, L.I., Ivanov, M.V., Lobas, A.A., and Gorshkov, M.V. (2017). Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. J. Proteome Res. *16*, 393–397.
55. Zhao, Y., and Jensen, O.N. (2009). Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. Proteomics *9*, 4632–4641.
56. Lange, E., Tautenhahn, R., Neumann, S., and Gröpl, C. (2008). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics *9*, 375.
57. Röst, H.L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B.C., Gillet, L., Testa, G., Malmström, L., et al. (2016). TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat. Methods *13*, 777–783.
58. Bro, R., Andersson, C.A., and Kiers, H.A.L. (1999). PARAFAC2—part II. Modeling chromatographic data with retention time shifts. J. Chemometrics: A J. Chemometrics Soc. *13*, 295–309.
59. Moseley, M.A., Hughes, C.J., Juvvadi, P.R., Soderblom, E.J., Lennon, S., Perkins, S.R., Thompson, J.W., Steinbach, W.J., Geromanos, S.J., Wildgoose, J., et al. (2018). Scanning quadrupole data-independent acquisition, part A: qualitative and quantitative characterization. J. Proteome Res. *17*, 770–779.
60. Messner, C.B., Demichev, V., Bloomfield, N., White, M., Kreidl, M., Ivosev, G., Wasim, F., Zelezniak, A., Lilley, K.S., Tate, S., et al. (2020). Scanning SWATH acquisition enables high-throughput proteomics with chromatographic gradients as fast as 30 seconds. bioRxiv. https://doi.org/10.1101/656793.
61. Zhu, X., Chen, Y., and Subramanian, R. (2014). Comparison of information-dependent acquisition, SWATH, and MSAll techniques in metabolite identification study employing ultrahigh-performance liquid chromatography–quadrupole time-of-flight mass spectrometry. Anal. Chem. *86*, 1202–1209.
62. Kossaifi, J., Panagakis, Y., Anandkumar, A., and Pantic, M. (2019). TensorLy: tensor learning in python. J. Mach. Learn. Res. *20*. https://www.jmlr.org/papers/volume20/18-277/18-277.pdf.
63. d'Atri, V., Causon, T., Hernandez-Alba, O., Mutabazi, A., Veuthey, J.-L., Cianferani, S., and Guillarme, D. (2018). Adding a new separation dimension to MS and LC–MS: what is the utility of ion mobility spectrometry? J. Sep. Sci. *41*, 20–67.
64. Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., and Engelsen, S.B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. *54*, 413–419.
65. Goloborodko, A.A., Levitsky, L.I., Ivanov, M.V., and Gorshkov, M.V. (2013). Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. J. Am. Soc. Mass Spectrom. *24*, 301–304.
66. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.
67. Bro, R., and Sidiropoulos, N.D. (1998). Least squares algorithms under unimodality and non-negativity constraints. J. Chemometrics *12*, 223–247.
68. Du, P., Kibbe, W.A., and Lin, S.M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics *22*, 2059–2065.
69. Virtanen, P., SciPy 1. 0 Contributors, Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.