Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes

Isabell Bludau1,2,8, Moritz Heusel1,3,8, Max Frank1,4, George Rosenberger1,5, Robin Hafen1, Amir Banaei-Esfahani1, Audrey van Drogen1, Ben C. Collins1,6, Matthias Gstaiger1 and Ruedi Aebersold1,7

Most catalytic, structural and regulatory functions of the cell are carried out by functional modules, typically complexes containing or consisting of proteins. The composition and abundance of these complexes and the quantitative distribution of specific proteins across different modules are therefore of major significance in basic and translational biology. However, detection and quantification of protein complexes on a proteome-wide scale is technically challenging. We have recently extended the targeted proteomics rationale to the level of native protein complex analysis (complex-centric proteome profiling). The complex-centric workflow described herein consists of size exclusion chromatography (SEC) to fractionate native protein complexes, data-independent acquisition mass spectrometry to precisely quantify the proteins in each SEC fraction based on a set of proteotypic peptides and targeted, complex-centric analysis where prior information from generic protein interaction maps is used to detect and quantify protein complexes with high selectivity and statistical error control via the computational framework CCprofler (https://github.com/CCprofler/CCprofler). Complex-centric proteome profiling captures most proteins in complex-assembled state and reveals their organization into hundreds of complexes and complex variants observable in a given cellular state. The protocol is applicable to cultured cells and can potentially also be adapted to primary tissue and does not require any genetic engineering of the respective sample sources. At present, it requires ~8 d of wet-laboratory work, 15 d of mass spectrometry measurement time and 7 d of computational analysis.

Introduction

Proteins are major effectors and regulators of biological processes and can elicit or participate in multiple functions depending on their interaction with other proteins; most catalytic, regulatory and structural functions are carried out by protein complexes. Therefore, it is of central and general interest in basic and translational biology to identify protein complexes in biological samples and to detect changes in their composition and/or abundance across different samples and states. Over the last decades, different strategies to systematically study protein-protein interactions (PPIs) and protein complexes have been developed. With increasing technical abilities in the area of mass spectrometry (MS)–based proteomics, high-throughput methodologies for detecting PPIs by affinity-purification coupled to MS (AP-MS) have emerged as gold-standard techniques for mapping large PPI interaction networks. However, the need to perform multiple reciprocal pull-downs, requiring either genetic engineering or availability of multiple specific antibodies, makes the AP-MS approach limited in its capability to comprehensively study the protein interactome across multiple conditions and to reliably detect in parallel concurrent changes in multiple complexes in a set of samples. Other techniques, such as BioID, are based on proximity labeling and obtain information about proteins that are in close spatial proximity inside the cell but that do not necessarily interact directly. Another recently developed technique couples native protein co-fractionation of complexes with MS followed by correlative analysis of chromatographic profiles to infer protein-protein interactions

1Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Zurich, Switzerland. 2Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. 3Division of Infection Medicine (BMC), Department of Clinical Sciences, Lund University, Lund, Sweden. 4European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. 5Columbia University, New York, NY, USA. 6School of Biological Sciences, Queen’s University of Belfast, Belfast, UK. 7Faculty of Science, University of Zurich, Zurich, Switzerland. 8These authors contributed equally: Isabell Bludau, Moritz Heusel. ✉E-mail: matthias.gstaiger@imsb.biol.ethz.ch; aebersold@imsb.biol.ethz.ch
(co-fractionation–MS (CoFrac-MS) or protein correlation profiling–MS)\textsuperscript{6–13}. Here, a single fractionation experiment in principle captures information on thousands of PPIs and hundreds of protein complexes in parallel and substantially increases the throughput for screening protein complexes.

Previous applications of CoFrac-MS strategies already provided interesting novel insights into the protein complex landscape across different organisms\textsuperscript{11} and molecular perturbations\textsuperscript{13}. They also showed that the CoFrac-MS approach faces the general limitation that the cumulative number of proteins detected across the fractions exceeds the number of fractions and by almost two orders of magnitude. Furthermore, the methods remained limited with regard to the resolution of chromatographic separations, the accuracy and consistency of mass spectrometric protein measurements and specificity in the inference of protein-protein interactions and the composition of protein complexes from the highly convoluted data\textsuperscript{4,13,15}.

With the goal to address these prevailing limitations and to assess proteome organization more quantitatively and precisely, we developed the complex-centric proteome-profiling strategy that builds on size exclusion chromatography (SEC)–based protein complex fractionation with increased chromatographic resolution, optimized peptide and protein quantification by data-independent acquisition (DIA) MS (also referred to as sequential window acquisition of all theoretical (SWATH)–MS) and the targeted detection of protein complexes at controlled error rates based on queries of putative protein complexes in complex-centric analysis\textsuperscript{16}. The core difference to other approaches lies in the concept of complex-centric analysis, which essentially changes the data analysis strategy of CoFrac-MS data from a discovery-based strategy to a multiple hypothesis testing–based strategy that uses prior information of protein complexes to suggest stable modules. In essence, the complex-centric strategy is an extension of the targeted proteomics rationale (Box 1) from the level of protein analysis to the level of protein complex analysis. The targeted query of protein complexes in the co-fractionation data generated by DIA/SWATH-MS improves selectivity compared to other, discovery-based data analysis approaches that generally focus on the detection of pairwise interactions from which complex compositions are then inferred (interaction centric\textsuperscript{15,17}). In addition to accurate detection, complex-centric analysis also extracts quantitative information on the composition and abundance of the cellular module(s) in which each detected protein participates. Complex-centric proteome profiling thus supports the parallel detection and quantification of hundreds of protein complexes at unprecedented resolution and bears significant potential to discover novel aspects of modular proteome function in diverse biological processes. The results of complex-centric proteome profiling can subsequently be leveraged to select protein complexes for further downstream characterization, for example, by validating the presence of unexpected complex variants (e.g., by reciprocal affinity purification or live cell imaging). The structure of protein complexes and their functional relevance can further be investigated by methods such as electron microscopy and functional assays.

Overview of the protocol

This protocol article provides step-by-step instructions to profile the higher-order complex assembly state of a given proteome via the complex-centric SEC-SWATH-MS workflow. The workflow consists of three main modules: (i) sample preparation via extraction of protein complexes under native conditions from a biological sample and their fractionation by SEC, (ii) data acquisition by bottom-up proteomics analysis of all collected fractions by DIA-MS combined with targeted, peptide-centric analysis (SWATH-MS) and (iii) computational inference of the overall proteome assembly state and detection of specific protein complexes by targeted, complex-centric analysis within the R software package CCprofiler.

The protocol takes as input a biological sample, exemplified by \(7\times 7\) HEK293 cells, as well as prior information on the tested protein interactions and/or complexes formed within the proteome of interest. The protocol produces as output: (i) a quantitative assessment of the global proteome assembly state, (ii) a quantitative assessment indicating how each protein partitions into a certain number of SEC-resolvable, distinct protein assemblies and (iii) a quantitative assessment of the interactions and protein complexes in the given proteome in the biological state tested. Figure 1 schematically summarizes the overall complex-centric proteome-profiling workflow by SEC-SWATH-MS.

The key requirements to successfully perform an SEC-SWATH-MS experiment and to analyze the resulting data are (i) the availability of biological specimens from which intact native complexes can be isolated, (ii) HPLC and DIA-enabled mass spectrometric equipment and (iii) computing infrastructure for data storage, peptide-centric SWATH-MS data analysis and complex-centric
SEC-SWATH-MS data analysis in the R environment. The built-to-task algorithms for complex-centric analysis are implemented in the R package CCprofi ler. Familiarity with the R programming language is not required for the use of the canonical complex-centric workflow but highly recommended if deviations from the canonical workflow as presented herein are expected.

Applications of the method
The key feature of complex-centric proteome profiling by SEC-SWATH-MS is its ability to assign separated proteins to hundreds of protein complexes in parallel from the same sample and to determine the relative distribution of protein mass across different assembly states. These analyses provide results at a controlled error rate and at sub-complex resolution. Preliminary results show that the analysis strategy can be extended to quantitatively compare protein complex abundance and composition and the distribution of proteins into alternative assembly states across different biological conditions, which is the focus of ongoing work.

Compared to other co-fractionation methods, the complex-centric workflow supports the rapid and error-controlled detection of hundreds of protein complexes from a single dimension of chromatographic fractionation and is supported by key improvements in quantitative profiling and data analysis. The workflow extends the targeted proteomics rationale as initially implemented by selected reaction monitoring and, more recently, by DIA/SWATH-MS toward the detection of protein complexes.
complexes from CoFrac-MS data. Specifically, the presence of specific protein complexes in the sample is tested by means of targeted queries of database-curated complexes or other sources providing prior information about the protein complexes in question. We could demonstrate that the targeted analysis rationale improves the overall selectivity of detecting reference complexes to a level comparable to that of a state-of-the-art CoFrac-MS workflow based on the analysis of a 12-fold higher number of fractions generated by multidimensional biochemical separations and binary interaction-centric data processing\textsuperscript{10,11,16}. In addition to indicating complex composition, the complex-centric workflow also provides estimates on the relative abundance of the different protein subunits in the detected complexes and the complex as a whole. Furthermore, the relative distribution of specific proteins across multiple complex assemblies can be quantified. These types of information are expected to closely correlate with the functional state of the respective protein modules, based on the principle that biochemical activities of a protein complex substantially depend on its subunit composition, topology and overall structure and that alterations of these parameters also alter activity and function\textsuperscript{19}. 

---

**Fig. 1 | Schematic overview of the complex-centric proteome-profiling workflow.** The workflow consists of three main modules: (i) extraction of protein complexes from a biological sample and fractionation by SEC, (ii) bottom-up proteomics analysis of all sampled fractions by DIA-MS combined with targeted, peptide-centric analysis (SWATH-MS) and (iii) inference of proteome assembly state and detection of specific protein complexes by targeted, complex-centric analysis within the R software package CCprofiler. The protocol takes as input a biological sample and requires some standard proteins for MW calibration, a spectral library for peptide-centric SWATH-MS data analysis and prior protein connectivity information. Outputs of the workflow are: (i) a quantitative assessment of the overall global proteome assembly state of the proteome analyzed, (ii) a quantitative assessment of how each protein partitions into a certain number of SEC-resolvable distinct protein assembly states and (iii) a quantitative assessment of the protein complexes and sub-complexes in the given proteome at its current biological state.
While the first exemplary application of the protocol was shown for the HEK293 cell line\(^{16}\), the complex-centric methodology can be readily adapted to different sample types that are compatible with mild lysis and the extraction of sufficient amounts of native protein complexes for SEC-SWATH-MS analysis (currently $\geq 1$ mg of total protein). The complex-centric proteome-profiling strategy is thus expected to enable the assessment of proteome organization and assembly states of specific protein complexes across a wide range of experimental model systems and perturbations\(^{18}\), potentially also including clinical specimens (the subject of ongoing work). Besides studying essential biological processes, such as the impact of growth factor treatment or cell differentiation, it can be envisaged that SEC-SWATH-MS can also be employed toward the characterization of drug mode of action with respect to protein complex assembly states or, in fact, for biomarker discovery. For example, complex-centric profiling might capture off-target effects that do not affect the level of transcripts or proteins but that alter cellular functions by altering the assembly state of specific protein complexes. Application of the complex-centric proteome profiling workflow to different experimental systems or across perturbations bears profound potential to provide novel insights into the interplay between proteome organization and function. A core focus of future work will therefore be to further improve method throughput to enable larger comparative or longitudinal studies, for example, by high-throughput on-line reversed-phase liquid chromatography\(^{20}\). Furthermore, we envisage the extension to additional fractionation techniques and the extension of the computational framework CCprofi ler to quantitatively compare protein complex assemblies across conditions both on the level of protein complex assembly states and protein complex composition.

The overall performance of the presented workflow is tightly linked to chromatographic resolution, the accuracy of mass spectrometric readout and the selectivity of complex-centric analysis. We thus recommend applying all three modules of the workflow as presented herein. However, depending on the specific biological question at hand and the available technology and expertise in a research laboratory, we anticipate that individual modules of the workflow can be replaced by alternative approaches or adapted to specific conditions. For example, protein complexes could be fractionated by other techniques than SEC, such as density gradient centrifugation, blue native poly-acrylamide gel electrophoresis or ion-exchange chromatography, as has been employed in other studies\(^{10,11,21}\). It should be noted, however, that the fractionation technique employed will determine key attributes such as resolution and observability of certain protein complexes that may impair workflow performance, especially in cases where no specific size information is obtained, as is the case in ion-exchange chromatography. Available size estimates of the detected protein complexes improve the selectivity of the canonical SEC-SWATH-MS workflow. Furthermore, as an alternative to the label-free DIA method used here, other methods to determine the abundance of proteins in consecutive fractions can be used. These include MS1-based or spectral counting–based analysis of data acquired by data-dependent acquisition (DDA). However, in our primary research paper, we showed that these methods have a substantially lower performance compared to the complex-centric workflow on a 5600+ MS platform (Fig. 2b, c in ref. \(^{16}\)). To compare multiple biological conditions and to eliminate potential variability of SEC fractionation, stable isotope labeling using amino acids in cell culture (SILAC)\(^{22}\) may be applicable but was as yet not tested in combination with the complex-centric profiling strategy. DDA with SILAC-based quantification was successfully employed in other co-fractionation–based studies\(^{8,10,14}\) and might provide quantitative data of sufficient quality to support complex-centric analysis. Similarly, multiplexing the analysis of multiple chromatographic fractions by isobaric labeling has not been tested. We anticipate that the performance of the complex-centric method will depend not only on the acquisition strategy but also on the liquid chromatography and tandem mass spectrometry (LC-MS/MS) platform employed. Finally, if the discovery of novel protein complexes is of interest, the data analyzed with regard to quantitative assembly states of previously known protein complexes via CCprofi ler can alternatively, or in addition, be employed to predict novel complexes via analysis tools such as PrInCe\(^{15}\) or EPIC\(^{23}\). This might be especially useful in cases where no or only limited prior knowledge on protein complexes is available for the organism analyzed.

**Experimental design**

Complex-centric proteome profiling by SEC-SWATH-MS was designed to maximize the resolution and depth of biological insights obtainable from a single protein co-fractionation experiment. This is achieved by extending concepts from the analysis of peptides via targeted proteomics to the level of protein complexes, thus implementing a targeted data analysis strategy. The workflow extracts quantitative information on the observed assembly states of predefined protein complexes with high
selectivity. The workflow consists of three main modules, each of which was optimized to address the prevailing limitations of previous co-fractionation workflows.

(1) Extraction of protein complexes from a biological sample and fractionation by SEC
To minimize complex disassembly under the diluted conditions present after cell lysis, all processing steps up to the collection of fractions are carried out rapidly, at minimal dilution and below 4 °C. To improve on protein complex resolution and to produce strongly correlated, sharp peak signals for the detected proteins, SEC is performed with high-resolution stationary phase material that offers an optimal tradeoff between resolution and fractionation range. A preferred column material combines 3-μm particle size and 500-Å pore diameters.

Protein complexes are isolated using a mild detergent, nonidet-P40, that has been shown to support the extraction and maintenance of integrity even of relatively labile protein complexes. To minimize the artifactual disruption of native protein complexes, cells are snap-frozen in liquid nitrogen before immediate processing. The cells are lysed by freeze-thawing into a low volume of 0.5% (vol/vol) detergent-containing lysis buffer supplemented with protease- and phosphatase inhibitors. These precautions are taken to minimize artifactual proteome rearrangements that do not reflect the true biological state of the system. After proteome extraction, lysates are cleared by ultracentrifugation. This step is pivotal to remove cellular components interfering with successful SWATH-MS analysis. To reduce detergent levels and to concentrate the sample for SEC fractionation, the buffer is exchanged over a 30-kDa molecular weight (MW) cut-off membrane in multiple dilution steps to minimize dilution-induced complex dissociation.

To monitor column performance and calibrate the apparent MW scale per each fraction, a five-protein standard sample is analyzed before and after the preparative protein complex fractionations required for an experiment or study (Fig. 2a). The elution profiles of standard proteins are further used to determine the fraction-collection scheme. We generally aim at collecting 7-8 independent MS measurements (fractions) across each eluting peak, a measure that has proven optimal in targeted proteomics applications aimed at the accurate quantification of peptides based on reconstruction of fragment ion chromatogram peak groups along peptide elution from on-line reversed-phase chromatography.

An aliquot of the unfractionated sample should be included for DIA/SWATH-MS analysis to obtain a ‘master’ sample and SWATH-MS dataset in which most signals observable across the individual fractions are represented. This dataset is used to optimize peptide-centric analysis, specifically to align parameters for statistical scoring of peptide fragment ion peak group signals across the full SEC-SWATH-MS study.

(2) Bottom-up proteomics analysis and peptide-centric analysis (SWATH-MS)
To improve the accuracy and consistency of mass spectrometric peptide signals along the SEC axis, DIA-based SWATH-MS is employed. Before sample acquisition, proteins contained in each collected SEC fraction are denatured by heating in 1% sodium de-oxoycholate (wt/vol), reduced, alkylated and digested with trypsin overnight. Compared to a range of alternative sample-workup regimens tested, the protocol described below provided the most robust results. The tryptic peptides are cleaned up by binding to a C18 resin and subsequent one-step elution. They are then dried and resuspended in nanoLC buffer A, supplemented with internal standard peptides to align peptide retention times in C18 chromatography, essentially as described. The inclusion of the reference peptides is especially important for peptide-centric, spectral library–based data analysis (see below).

To maintain the ability to quantitatively compare peptide/protein abundances across individual fractions of the recorded SEC-SWATH-MS maps, we recommend the injection of equal volumes for each SEC fraction for analysis by SWATH-MS. We recommend determining the injection volume for a specific SEC separation based on test injections of the first two SEC fractions that produce the highest absorbance at 280 nm in UV-visible profiling during SEC fractionation. The injection volume is then selected to maximize the amount of peptides injected per MS run without exceeding the upper sample limit tolerated by the mass spectrometer (Fig. 2b). This upper limit is determined by detector saturation. For example, for a 5600+ instrument operated in SWATH mode using 64 variable isolation windows, the maximal total ion current (TIC) is $1 \times 10^8$ (here, fraction 55; also see Fig. 2c). As a consequence of keeping the volume injected from each SEC fraction constant, fractions containing smaller amounts of total protein will generate SWATH maps that top out significantly below the maximally tolerated ion current. The identity of the most highly concentrated fractions and
optimal injection volume may depend on sample type, SEC setup and MS platform used. We recommend including an aliquot of the unfraccionated sample in the final acquisition queue.

While other MS platforms and MS acquisition schemes may in principle also be compatible with complex-centric profiling, we advise employing a DIA/SWATH-MS based approach to achieve highly consistent detection and accurate quantification of a consistent set of peptides across the SEC dimension. The quality of the peptide-level data is critical for the performance of complex-centric proteome-profiling analysis in CCprofiler, as shown in our original publication\textsuperscript{16}. The workflow described was successfully implemented using the ABSciex 5600+ or 6600+ MS platforms in combination with the 64 variable window acquisition scheme (Tables 1, 2 and 3)\textsuperscript{16,30}. Other quantitative DIA workflows as implemented on alternative mass spectrometric platforms may also produce data of sufficient quality for effective complex-centric proteome profiling\textsuperscript{31–33}. We recommend acquisition of SWATH-MS maps along the SEC fractions in 120-min gradients, but recent studies have shown promising results when employing shorter (e.g., 60 min) nanoLC or super-short LC gradients (20 min)\textsuperscript{20}.

The interpretation of DIA/SWATH-MS data is most effective in conjunction with a reference spectral library that contains information about peptide elution, MS fragmentation and other pertinent properties. These parameters then support the detection and quantification of the respective peptides by targeted data analysis. Whereas samples from less intensely studied species may require the generation of a customized spectral library from DDA data acquired on the same sample\textsuperscript{28}, for more extensively studied species, including humans, mice, yeast and zebrafish, large-scale libraries are
For applications involving human samples, we recommend using the combined human assay library (CAL), which combines peptide query parameters from over 300 DDA-MS injections across diverse human cell and tissue types. The CAL also contains spectral information from our SEC-SWATH-MS analysis of the HEK293 proteome and is expected to be sufficiently representative also for future studies of human samples via SEC-SWATH-MS. In the course of developing the SEC-SWATH-MS methodology, the use of a sample-specific library versus the CAL was evaluated. The results showed benefits of the CAL with respect to increased sequence coverage and increased number of peptides per protein. If a given research question depends on a sample-specific spectral library, we recommend DDA acquisition side by side with the DIA/SWATH-MS measurements and library construction as described previously.

Once all DIA/SWATH-MS data are acquired, data are converted to mzXML format for open-source peptide-centric analysis via the OpenSWATH workflow, which consists of OpenSWATH, PyProphet and TRIC analysis.

| Table 1 | LC parameters |
|---------|---------------|
| Parameter | Setting/value |
| Injection volume | 4 µl |
| Flow rate | 0.3 µl/min |
| LC gradient | Time (min): nanoLC buffer B percentage (%B): |
| 0:00 | 2 |
| 120:00 | 35 |
| 125:00 | 90 |
| 121:00 | 90 |
| 126:00 | 2 |
| 135:00 | 2 |

| Table 2 | MS acquisition: general parameters |
|---------|------------------------------------|
| Parameter | Setting/value | Comment |
| MS1 mass range | 360–1,460 m/z |
| MS1 accumulation time | 250 ms | Long fill time for usability in MS1 scoring in OSW + PyProphet |
| MS2 mass range | 300–2,000 m/z |
| MS2 accumulation time | 50 ms |
| Curtain gas flow rate (CUR) | 28.000 |
| Ion source gas 1 (GS1) | 15.000 |
| Ion source gas 2 (GS2) | 0.000 |
| Interface heater temperature (IHT) | 150.000 |
| IonSpray voltage (ISVF) | 2,600.000 |
| Collision energy spread (CES) | 15.000 |
| Declustering potential (DP) | 100.000 |
| IDIx | 0.000 |
| IDUx | 5.000 |
| Ion release delay (IRD) | 66.633 |
| Ion release delay x (IRDx) | 28,287.602 |
| Ion release width (IRW) | 24.917 |
| Ion release width x (IRWx) | 24,917.307 |
| IWIx | 0.000 |
| IWUx | 5.000 |
| Scan          | Start mass (m/z) | End mass (m/z) | Collision energy (CE) | Radio frequency amplitude (XA1) |
|--------------|-----------------|---------------|-----------------------|---------------------------------|
| TOF MS1      | 350             | 1,460         | 10                    | 166.3                           |
| SWATH exp. 1 | 399.5           | 408.2         | 22.1                  | 134.2                           |
| SWATH exp. 2 | 407.2           | 415.8         | 22.6                  | 135.4                           |
| SWATH exp. 3 | 414.8           | 422.7         | 23                    | 136.6                           |
| SWATH exp. 4 | 421.7           | 429.7         | 23.5                  | 137.7                           |
| SWATH exp. 5 | 428.7           | 437.3         | 23.9                  | 138.9                           |
| SWATH exp. 6 | 436.3           | 444.8         | 24.4                  | 140.1                           |
| SWATH exp. 7 | 443.8           | 451.7         | 24.8                  | 141.3                           |
| SWATH exp. 8 | 450.7           | 458.7         | 25.3                  | 142.4                           |
| SWATH exp. 9 | 457.7           | 466.7         | 25.7                  | 143.5                           |
| SWATH exp. 10| 465.7           | 473.4         | 26.2                  | 144.8                           |
| SWATH exp. 11| 472.4           | 478.3         | 26.6                  | 145.8                           |
| SWATH exp. 12| 477.3           | 485.4         | 26.9                  | 146.6                           |
| SWATH exp. 13| 484.4           | 491.2         | 27.4                  | 147.7                           |
| SWATH exp. 14| 490.2           | 497.7         | 27.7                  | 148.7                           |
| SWATH exp. 15| 496.7           | 504.3         | 28.1                  | 149.7                           |
| SWATH exp. 16| 503.3           | 511.2         | 28.6                  | 150.8                           |
| SWATH exp. 17| 510.2           | 518.2         | 29                    | 151.9                           |
| SWATH exp. 18| 517.2           | 525.3         | 29.4                  | 153                             |
| SWATH exp. 19| 524.3           | 533.3         | 29.9                  | 154.2                           |
| SWATH exp. 20| 532.3           | 540.3         | 30.4                  | 155.4                           |
| SWATH exp. 21| 539.3           | 546.8         | 30.8                  | 156.5                           |
| SWATH exp. 22| 545.8           | 554.5         | 31.2                  | 157.6                           |
| SWATH exp. 23| 553.5           | 561.8         | 31.7                  | 158.8                           |
| SWATH exp. 24| 560.8           | 568.3         | 32.1                  | 160                             |
| SWATH exp. 25| 567.3           | 575.7         | 32.6                  | 161                             |
| SWATH exp. 26| 574.7           | 582.3         | 33                    | 162.2                           |
| SWATH exp. 27| 581.3           | 588.8         | 33.4                  | 163.3                           |
| SWATH exp. 28| 587.8           | 595.8         | 33.8                  | 164.3                           |
| SWATH exp. 29| 594.8           | 601.8         | 34.3                  | 165.4                           |
| SWATH exp. 30| 600.8           | 608.9         | 34.7                  | 166.4                           |
| SWATH exp. 31| 607.9           | 616.9         | 35.1                  | 167.6                           |
| SWATH exp. 32| 615.9           | 624.8         | 35.6                  | 168.8                           |
| SWATH exp. 33| 623.8           | 632.2         | 36.1                  | 170.1                           |
| SWATH exp. 34| 631.2           | 640.8         | 36.6                  | 171.3                           |
| SWATH exp. 35| 639.8           | 647.9         | 37.1                  | 172.6                           |
| SWATH exp. 36| 646.9           | 654.8         | 37.5                  | 173.8                           |
| SWATH exp. 37| 653.8           | 661.5         | 38                    | 174.9                           |
| SWATH exp. 38| 660.5           | 670.3         | 39                    | 176                             |
| SWATH exp. 39| 669.3           | 678.8         | 39.5                  | 177.4                           |
| SWATH exp. 40| 677.8           | 687.8         | 40.1                  | 178.8                           |
| SWATH exp. 41| 686.8           | 696.9         | 40.6                  | 180.2                           |
| SWATH exp. 42| 695.9           | 706.9         | 41.2                  | 181.7                           |
| SWATH exp. 43| 705.9           | 715.9         | 41.8                  | 183.3                           |
| SWATH exp. 44| 714.9           | 726.2         | 42.5                  | 184.8                           |
| SWATH exp. 45| 725.2           | 737.4         | 43.2                  | 186.4                           |
| SWATH exp. 46| 736.4           | 746.6         | 43.8                  | 188.2                           |
| SWATH exp. 47| 745.6           | 757.5         | 44.4                  | 189.7                           |
| SWATH exp. 48| 756.5           | 767.9         | 45.1                  | 191.4                           |
| SWATH exp. 49| 766.9           | 779.5         | 45.8                  | 193.1                           |
| SWATH exp. 50| 778.5           | 792.9         | 45.8                  | 195                             |
(3) Analysis of the global proteome assembly state and specific protein complexes by targeted, complex-centric analysis via CCprofi ler

The overall goal of the complex-centric workflow is the inference of the global proteome assembly state and the detection of specific protein complexes from the measured SEC protein profiles. We developed CCprofi ler, an algorithm and computational framework to support these analyses. After initial data preparation and import, CCprofi ler performs three main steps: (i) data pre-processing and quality control, (ii) protein-centric analysis and (iii) complex-centric analysis. In the following, the rationale behind each of these steps is explained.

Data preparation and import

Before analysis in CCprofi ler, all necessary input data first need to be formatted and imported into R according to the CCprofi ler guidelines (Table 4). The main input for CCprofi ler is the quantitative peptide-level data, which need to be imported into a ‘traces object’. A ‘traces object’ is a customized data format of the CCprofi ler package. Traces objects can store information about quantitative peptide or protein profiles and can further include additional annotations of the measured peptides, proteins and fractions. For direct data import of the TRIC output, the importFromOpenSWATH function can be used. It removes non-proteotypic evidence and sums precursor signals per peptide to generate peptide-level quantitative information. If an alternative DIA (or DDA) analysis platform is selected, non-proteotypic peptides should be removed manually if necessary, and precursor signals should be summarized to unique peptide quantitative information. The quantitative peptide matrices generated thereby can also be imported into traces objects via the importPCPdata function. In addition to (i) the quantitative peptide-level information, CCprofi ler requires (ii) a fraction annotation table that maps each MS run to a given fraction; (iii) an MW calibration table generated by measuring the apex fractions (i.e., the fractions corresponding to the highest intensity of a protein’s elution profile) of an external standard set of reference proteins fractionated on the same SEC setup, used to establish a log-linear relationship between the SEC fractions and their apparent MW (Figs. 2a and 3a); (iv) a trace annotation table containing information from UniProt (https://www.uniprot.org/) that is used to annotate proteins with their associated gene names and monomeric MWs; and (v) prior protein connectivity information in the form of defined protein complexes (e.g., annotated in CORUM44) or binary interaction networks, such as StringDB45,46 or BioPlex2,3. For each new dataset being analyzed by complex-centric analysis, we suggest performing an initial analysis starting from a set of well-defined protein complexes. The CORUM database provides this information for several species including humans, mice, rats and generic mammalian assemblies44. This will enable researchers to get a first impression of the quality and protein complex recovery achievable from the given SEC-SWATH-MS experiment for most eukaryotic systems. Alternatively, complex queries can be generated based on interaction networks derived from different experimental techniques such as functional complementation47,48 or AP-MS (e.g., BioPlex network2,3), or from networks integrating

| Scan       | Start mass (m/z) | End mass (m/z) | Collision energy (CE) | Radio frequency amplitude (XA1) |
|------------|------------------|----------------|-----------------------|---------------------------------|
| SWATH exp. 51 | 791.9           | 807            | 46.7                  | 197.2                           |
| SWATH exp. 52 | 806             | 820            | 47.6                  | 199.4                           |
| SWATH exp. 53 | 819             | 834.2          | 48.4                  | 201.5                           |
| SWATH exp. 54 | 833.2           | 849.4          | 49.3                  | 203.8                           |
| SWATH exp. 55 | 848.4           | 866            | 50.2                  | 206.3                           |
| SWATH exp. 56 | 865             | 884.4          | 51.3                  | 209                             |
| SWATH exp. 57 | 883.4           | 899.9          | 52.4                  | 211.9                           |
| SWATH exp. 58 | 898.9           | 919            | 53.4                  | 214.5                           |
| SWATH exp. 59 | 918             | 942.1          | 54.7                  | 217.7                           |
| SWATH exp. 60 | 941.1           | 971.6          | 56.2                  | 221.6                           |
| SWATH exp. 61 | 970.6           | 1,006          | 58.1                  | 226.4                           |
| SWATH exp. 62 | 1,005           | 1,053          | 60.4                  | 232.3                           |
| SWATH exp. 63 | 1,052           | 1,110.6        | 63.5                  | 240.2                           |
| SWATH exp. 64 | 1,109.6         | 1,200.5        | 67.5                  | 250.4                           |
multiple layers of physical and functional interaction evidence (e.g., StringDB\textsuperscript{45,46}). Some of these resources of prior protein connectivity information are available for different species\textsuperscript{49–51}. Notably, StringDB covers multiple species with experimental evidence and extrapolates proteome connectivity through orthology mapping to as many as 5,090 species. In addition, complex queries can be formulated manually to test for the presence of suspected assemblies in the sample. The scores of such ad hoc queries are evaluated in the context of other, database-derived queries and interpreted in the context of a global error model. Thus, the complex-centric analysis approach is expected to be applicable to a wide range of different organisms and sample types and will continuously improve as the reference interaction knowledge is refined over time. A summary of the required input data and their format for a successful CCprofiler analysis are presented in Table 4.

### Data pre-processing and quality control

To achieve optimal data quality and sensitivity in complex-centric profiling, CCprofiler includes several functions to increase data completeness and to filter peptides for detection and quantification consistency.

Once a traces object is imported, missing values can be detected based on a user-defined criterion. In most proteomics pipelines, zero intensity values indicate either that the signal is ‘missing at random’ (i.e., no detection due to technical reasons such as interference from other peptides) or ‘missing not at random’ (i.e., no detection due to cellular concentrations below the detection limit). We suggest that a zero value is considered as ‘missing at random’ in case a quantitative (non-zero) signal has been detected in both the two fractions preceding and following the fraction in question. The detected ‘missing at random’ values are subsequently imputed by a spline fit across the fractionation dimension. In our originally published analysis workflow\textsuperscript{16}, we did not perform a missing value imputation. Generally, when applying a SWATH-MS–based workflow, only a few missing values are expected. However, sample loss (e.g., of an entire SEC fraction) sometimes cannot be avoided, thus generating missing values for transparent technical reasons. To still enable a robust and sensitive analysis with CCprofiler in such cases, we implemented the missing value detection and

### Table 4 | Summary of CCprofiler input data

| Input table                                      | Required columns                                                                 | Dependencies                                                                 |
|--------------------------------------------------|----------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| **Sample-specific data**                         |                                                                                  |                                                                               |
| Quantitative peptide-level data                  | TRIC output                                                                      | Required for each sample (unique condition and replicate)                    |
|                                                  | Wide format: protein_id, peptide_id, filename1, filename2,..., filenameN (values = MS intensity) |                                                                               |
|                                                  | Long format: protein_id, peptide_id, filename, intensity                         |                                                                               |
| Fraction annotation                               | filename, fraction_number                                                        | Required for each sample (unique condition and replicate)                    |
| MW calibration                                   | std_weights_kDa, std_elu_fractions                                               | Required for the assessment of the global proteome assembly state and for increased specificity of the complex-centric analysis module |
| **Generic data**                                 |                                                                                  |                                                                               |
| Trace annotation                                  | Requires one column that matches either peptide or protein ids in your traces object and additional columns with meta-information to be added to your traces object under $trace.annotation. A typical annotation table is the UniProtKB annotation on protein level, downloadable in .tsv format, including for example ‘Gene_names’ and protein ‘Mass’. | Required for the assessment of the global proteome assembly state |
| Reference database of protein complexes or binary PPIs | Protein complexes: any number of proteins supposedly forming a complex defined with columns: complex_id, complex_name, protein_id | Required only for complex-centric analysis module |
|                                                  | PPIs: Binary protein-protein interaction table with columns a,b and protein_id entries as employed in the traces_object |                                                                               |
imputation approach. While this does not make a large difference in the HEK293 SEC-SWATH-MS dataset presented in this paper, we specifically recommend this pre-processing step for datasets generated by DDA or if single fractions were lost during sample processing.

In a next step, peptides that have not been detected in more than N consecutive fractions (here, N = 2) are removed from the traces object. This effectively removes false-positive peptide detections from the dataset. Finally, we leverage the idea that multiple peptides originating from the same protein should display highly similar quantitative elution profiles along the chromatographic dimension, given that the proteins are presumably intact during separation and are only cleaved into peptides for LC-MS/MS analysis thereafter. For each peptide, the average pairwise correlation with the quantitative traces of its sibling peptides (i.e., peptides derived from the same protein) is calculated (Fig. 3b). Peptides below a minimum average sibling peptide correlation (SPC) cutoff can subsequently be removed. The rationale for this step is that outlier peptides and proteins with very heterogeneous quantitative peptide traces are excluded from further analysis. Given that decoy proteins are maintained in the upstream analysis, CCpro can automatically determine an SPC cutoff at which a user-defined criterion on the maximally acceptable global protein-level false discovery rate (FDR) is satisfied (Fig. 3c). FDR estimation can be fine-tuned by providing a prior probability on protein detectability in the form of a fraction of false targets (FFT). A conservative FFT can be estimated from the protein-level PyProphet scoring of the unfractonated SEC input sample. This is conservative, because we expect to see cumulatively more proteins in the SEC fractions than in the single unfractonated input sample. Alternatively to the automated SPC cutoff estimation, a user-provided SPC cutoff can be used for peptide filtering. Figure 3e shows the peptide-level quantitative profiles of the proteasome subunit alpha type-1 after pre-processing and filtering.

The peptide-level traces can subsequently be used for protein quantification. The protein-level profiles are inferred from summing the top N (we recommend N = 2) most-abundant and well-correlated peptides of a given protein. Only peptides that map uniquely to the protein of interest and in the context of the full-sequence database (proteotypic peptides) are used in the analysis. For the inference of protein complexes, the distribution of identified proteins across the SEC fractionation range is important. Supplementary Fig. 1a illustrates the percentage overlap of proteins detected in any fraction pair. While adjacent fractions share, on average, 92% of detected proteins, this number decreases to 56% at a distance of 10 fractions and levels off at 31% at a distance of 40 fractions.

The protein-level profiles can then be used to estimate the overall complex assembly state observed in the sample as a quality control to ensure the successful extraction and profiling of largely intact complexes (Fig. 3d). The protein-level profiles are also the input for complex-centric signal detection.

**Protein-centric signal detection of protein assembly states**

Protein-centric analysis aims to evaluate the number of distinct assembly states in which each protein is observed. We define an assembly state as a distinct SEC peak in which the protein in question is confidently detected. CCpro determines such distinct SEC peaks by detecting peptide co-elution peak groups along the chromatographic dimension. Each detected peak (‘protein feature’) represents the protein in a specific assembly state (i.e., monomeric or bound to different protein complexes) as inferred from the protein’s monomeric MW and external size calibration of the SEC fractionation. This analysis yields a fine-grained view of individual assembly states of each protein but also enables more global assessments of the overall degree of higher-order assembly observed from the biological sample.

The optimal parameters for the peak detection algorithm in CCpro depend on the chromatographic resolution and the quantitative accuracy of a given dataset. Therefore, CCpro features functions to automatically screen and identify optimal parameters based on a parameter grid search (Box 2 and Fig. 4a). The grid search screens different parameter combinations for their sensitivity to detect protein elution signals from sibling peptides (derived from the same protein) as opposed to nonsense signals among peptides sampled randomly from different proteins across the dataset. In our previously published analysis workflow\(^\text{16}\), parameter optimization was based on the detection of protein complex signals from protein traces. In contrast, we here present and suggest parameter optimization based on protein-level peak detection from peptide traces, because it is independent of the detectability of a certain set of protein complexes in the sample and because the chromatographic and quantitative properties are specific to the analyzed dataset and not the level of analysis (complex or protein elution).
Fig. 3 | Data pre-processing plots. a, MW calibration based on measured standard proteins and their MWs were as follows: thyroglobulin tetramer, 1,398 kDa; thyroglobulin dimer, 699 kDa; IgA, 300 kDa; IgG, 150 kDa; ovalbumin, 44 kDa; and myoglobin, 17 kDa. b, Distribution of SPCs for both target proteins (solid line) and decoy proteins (dashed line). c, Pseudo-ROC curves illustrating the effect of using the SPC to perform FDR filtering. d, Global statistics of protein signal attribution to assembled or monomeric state. Most detected protein mass (55%), as estimated by the total MS signal intensity, appears in assembled state in SEC-SWATH-MS. e, Elution profiles of all peptides detected for the proteasome subunit alpha type-1 protein. The red vertically dashed line indicates the expected monomer MW. The salmon-colored line indicates the selected cutoff for dividing the elution range into assembled (left) versus monomeric (right). The cutoff was set at greater than or equal to twice the expected monomer MW.
**Box 2 | Selection of appropriate feature finding parameters**

For the protein- and complex-centric signal detection to be successful, it is essential to select appropriate parameters. These include:

- **corr_cutoff**: minimum average Pearson correlation for a region to be selected by the sliding window algorithm for peak detection
- **window_size**: number of SEC fractions for the sliding window
- **smoothing_length**: number of fractions used for smoothing of the quantitative profiles for peak picking
- **rt_height**: maximum number of fractions for which detected peak signals are collapsed within one query

The optimal parameter set for signal detection depends critically on the chromatographic resolution of the applied co-fractionation strategy and the quantitative accuracy of the MS-based readout. Therefore, we designed a grid search strategy in CCprofiler to systematically screen for optimal parameters that most sensitively retrieve true positive co-elution signals in each individual dataset. However, it is essential to select reasonable search parameters at the beginning. Here, we discuss, based on our experience, how to select a reasonable set of parameters for the grid search.

The window_size, smoothing_length and rt_height parameters are conceptually linked, since they all depend on the expected peak width and resolution of chromatographic separation. To select a good range of window sizes, we suggest randomly plotting the elution profiles of several proteins and approximating the average number of fractions under the observable protein elution peaks. The selected window_size parameter should not be higher than the observed peak width. We suggest testing values between half and full peak width. The smoothing_length parameter can be tested in a similar range as the window_size parameter, selecting smaller rather than larger values.

The rt_height parameter determines the maximum offset between peak apex fractions to still be considered eluting in the same peak. The more sharp the detected peaks are and the more accurate the quantification is, the lower this parameter can be. We suggest choosing parameters between zero and maximally half the observed elution peak width.

The corr_cutoff represents the most critical parameter for the selectivity of peak detection. How high it can be set depends greatly on the quantitative accuracy and precision of the applied MS strategy. A good measure to estimate dataset quality is the overall ‘goodness of correlation’ observed among sibling peptides, which should, theoretically, correlate perfectly along the SEC dimension, since they were covalently bound as part of the same polypeptide chain during fractionation. Thus, the corr_cutoff parameter range worth screening can be inferred from the observed SPC density plot as produced during data pre-processing (Fig. 3). The minimal corr_cutoff to test should be the value for which most proteins have a good SPC (in our case, 0.8). Higher corr_cutoff parameters up to 0.95 could further be tested to potentially enable optimal selectivity.

Scoring and FDR estimation of the detected protein features are based on a target decoy strategy. Decoys are generated automatically, by randomly shuffling peptide-to-protein assignments. The co-elution scores calculated by CCprofiler (for more details, also see information available in the appendix of the original publication15) are converted into empirical P values and used for q value estimation52. To ensure that the decoy-based FDR estimation approach works correctly, the quantitative data need to be of sufficient quality, and a sufficient number of proteins need to be evaluated...
in parallel. We suggest that ≥500 proteins with high-quality SEC patterns that pass upstream consecutive identification and SPC-based filters should be used. To assess whether the FDR estimation worked correctly, it is crucial to manually inspect the $P$ value histogram generated by CCprofiler (Fig. 4b,c). There should be a high peak close to zero and a uniform distribution across all other
**Fig. 4** | Parameter selection and protein-centric analysis. a, Pseudo-ROC curves showing the number of estimated true positive protein features over increasing q value (-FDR) cutoffs for all tested parameter combinations. Here, each parameter set is colored according to the tested correlation cutoff. Corr, Pearson correlation. b, P value histogram for the protein-centric signal detection. c, Pseudo-ROC curve showing the number of estimated true positive protein features over increasing q value (-FDR) cutoffs. d, Histogram showing the number of proteins detected to elute in between one and seven distinct elution peaks. The pie chart illustrates that the majority of detected protein elution signals elute in the assembled MW range. e, Elution profiles of all peptides detected for the proteosome subunit alpha type-1 (PSA1) protein. The protein elution signals determined by CCprofiler are highlighted in gray shading; peak apexes (solid) and boundaries (dashed) are shown as gray vertical lines. The red vertically dashed line indicates the expected monomer MW. Max., maximum.

The confidently detected protein features each represent a unique protein assembly state, which allows the inference of several interesting biological conclusions. First, the distribution of protein features into signals likely representing monomeric or assembled forms of the protein can be used to draw conclusions about the overall assembly state of the cellular system tested (pie chart in Fig. 4d). Second, the distribution of proteins into a single or multiple distinct assembly states can reveal potential moonlighting of a protein across distinct protein assemblies in which it may assume alternative functional roles (bar chart in Fig. 4d). Protein feature signals do, however, not yet assess the exact interaction partners involved, which is a question that is assessed in the subsequent complex-centric signal detection module.

**Complex-centric detection of protein complex signals**

The heart of the CCprofiler software and the presented protocol is the complex-centric detection of protein complex signals. Here, prior protein connectivity information is used to query the SEC-SWATH-MS data for evidence of specific protein complexes or sub-complexes thereof in the biological sample. In analogy to peptide- and protein-centric analysis, complex-centric analysis detects protein co-elution peak groups along the chromatographic dimension based on a priori-defined protein complex queries and tests the hypothesis that the query complex is present in the sample. A single protein complex query frequently results in the detection of multiple distinct subunit co-elution signals (‘protein complex features’). Each of these signals corresponds to a different (sub-)version of the queried protein complex with distinct composition and/or stoichiometry, eluting at a distinct elution time. Depending on the completeness of prior information and MS observability (Fig. 5a), targeted, complex-centric analysis may or may not capture all involved protein subunits. In contrast to other analysis strategies of CoFrac-MS datasets that aim to predict new complexes¹⁷,²³, the complex-centric workflow aims to confidently detect a priori-defined protein complexes and variants thereof and to quantify the relative protein subunit distribution across these in the given biological state.

Prior protein connectivity information can be provided either directly in the form of the concrete composition of query complexes (‘complex queries’) or in the form of pair-wise PPI networks. An exemplary set of target complex queries are the complexes represented in CORUM⁴⁴. If a PPI network is provided, it first needs to be partitioned into smaller sets of interacting proteins, thereby generating concrete protein complex queries. This is achieved by simplistic network partitioning, selecting for each protein its immediate neighbors (degree = 1). Any redundant, fully overlapping complex queries are removed such that, per unique query complex, one decoy query complex is generated, and the assumptions for the error control strategy are met. However, we suggest keeping protein complex subset queries, such as a first complex query including subunits A, B and C and a second complex query including subunits A, B, C, D and E. The reason for also keeping the sub-complex query ABC is that the co-elution signal A-B-C will be scored preferably in the CCprofiler co-elution score when compared to the same signal scored as part of a larger complex query including subunits D and E. This is because scores are calculated in dependence of the number of proteins per query and the number of proteins observed co-eluting. Maintaining the subset hypothesis A-B-C will therefore increase the chance of successfully recovering and quantifying the A-B-C complex signal, particularly in the situation where subunits D and/or E do not partake in the observed protein complex and co-elution signal (for more details, also see the appendix on ‘Coelution score calculation and statistical FDR control’ in ref. ¹⁷). For cases in which both the A-B-C and A-B-C-D-E complex...
query point to the same protein complex signal in the SEC dimension, these detected features will finally be collapsed into one unique signal after the initial peak detection step (see below).

To enable an automated error-estimation of the complex-centric feature finding, CCprofiler employs a decoy-based FDR estimation strategy. A decoy complex query is generated for each target, thus resulting in the same size distribution of protein complex subunits for both targets and decoys. Decoys are generated by randomly assigning proteins to a specific decoy complex query. To ensure that the randomized protein-to-decoy-complex associations do not contain true interactions, we exclude direct interaction partners present in the target protein complex queries. This is achieved by selecting a minimum pairwise network distance N (here, N = 2). The target-decoy FDR control strategy of CCprofiler depends on a minimal number of target (and decoy) queries. The reasons are that (i) a sufficiently large binary interaction network needs to be available to randomly generate decoy complex queries and (ii) decoy-based FDR estimation is appropriate only for a representative number of detected target and decoy queries to ensure a stable P and q value estimation. We suggest a minimum number of 1,000 protein complex targets and decoys.

After the target and decoy protein complex queries are defined, complex-centric feature finding is performed. Here, protein traces are queried for the presence of local co-elution signals of the subunits specified in each of the target and decoy queries. Co-elution scores are calculated based on both the local correlation of the detected subunits across the SEC fractions and the fraction of correlating subunits relative to the total number of queried subunits. For the estimation and control of FDR, the most-complete co-elution signal for each protein complex query (i.e., the peak with the highest number of co-eluting protein subunits) is selected and used to convert co-elution scores into empirical P and q values. To confirm the correct operation of the error estimation and the fidelity of the results, it is important to manually inspect the P value histogram generated by CCprofiler (Fig. 5b, c). There should be a high peak close to zero and a uniform distribution across all other P values (also see Troubleshooting and refs. 53,54). To provide an estimate of the distribution of protein complexes across the SEC fractionation dimension, Supplementary Fig. 1b illustrates the percentage overlap between detected protein complexes in any fraction pair. While adjacent fractions share, on average, 90% of detected protein complexes, this number is reduced to 28% at a distance of 10 fractions and drops below 10% at a distance of >17 fractions. To interrogate potential protein complex assembly intermediates or other protein complex variants that are observable from a set of query subunit profiles additional to the best scoring peak group, secondary peak groups of each complex query are subsequently appended to the FDR filtered list of detected protein complex signals. For these assignments, a less-strict minimum local peak-correlation cutoff is selected manually specifically for the respective dataset.

The list of confidently detected protein complex features contains information not only about the presence and abundance of individual protein complexes, but further entails information about proteome modularity such as protein complex assembly intermediates and the relative abundance of the different protein subunits (also see Fig. 5d,e and Anticipated results).

Up to this point in the complex-centric analysis workflow, each protein complex signal detected by CCprofiler is directly linked to one specific protein complex query provided by the prior protein connectivity information. However, some of the subunits in each complex query might overlap with other complex queries. This redundancy in complex queries, in combination with the possibility of observing subsets/sub-complexes of these queries, can result in the reporting of multiple protein complex signals based on only one piece of experimental evidence (co-elution signal). For example, complex query A consists of subunits WXYZ, and complex query B consists of subunits VXYZ. If a co-elution signal among XYZ is detected in the data, it will, until this point, be reported for both complex queries A and B. Therefore, to retrieve truly unique signals, the co-elution...
signals need to be collapsed based on a strategy that considers (i) subunit composition and (ii) position in the chromatographic dimension (i.e., apparent MW of the protein complex). In the example above, signal collapsing will merge the redundant report of the XYZ signal derived from the two partially overlapping complex queries A and B into one unique reported protein complex signal (for more details, also see ref. 16). In case multiple signals with overlapping components are detected at the same elution fraction, unique composite signals will be generated, based on user-defined parameters on subunit overlap and proximity in elution fraction (e.g., signals containing the subunits XYZ, YZK and YZF all with apex fraction 32 will result in one unique reported signal XYZKF). It is important to note that the FDR estimation strategy in CCprofiler operates at the level of protein complex queries and does not propagate to the level of collapsed protein complex signals, and the precise error estimations among those collapsed results remain de facto unknown and the subject of future work.

Limitations
There are several limitations associated with the presented protocol. First, the workflow is optimized toward the analysis of soluble, cytosolic protein complexes that are extractable under native conditions and that remain stable during the multiple steps of the protocol through SEC separation.

Conceptually, the targeted, complex-centric analysis strategy is not designed to identify any novel protein complexes. Rather, it focuses on the detection and quantification of protein complexes annotated in public, generic protein interaction maps45,46. Therefore, the workflow is limited by the availability and coverage of such prior protein connectivity information. While the quality of the chosen prior interaction network naturally influences the results obtainable by complex-centric analysis, the co-elution signal-detection step and the FDR model in CCprofiler provide a good strategy to reduce the negative impact of false or inaccurate protein complex assignments represented in the prior network. Overall, the targeted approach is more sensitive and selective for protein complex queries with higher numbers (>3) of protein subunits, because random co-elution of these subunits becomes less likely16 compared to protein complexes consisting of a lower number of subunits. Even though complexes with completely new protein subunit composition cannot be detected with the described workflow at its current state, it nevertheless supports the identification of protein sub-complexes with partially unknown composition, as exemplified by the delineation of the mini-COP9 signalosome and different proteasome assembly intermediates16. In future work it is conceivable to further extend the targeted, complex-centric framework to further suggest additional subunits that are observed to co-elute with a given protein-complex signal. This would enable the de novo prediction of larger protein complex assemblies beyond prior knowledge.

In terms of quantification, complex-centric proteome profiling offers insights into: (i) the relative mass distribution of a specific protein between assembled and monomeric state, (ii) the relative mass distribution of a protein into multiple distinct assembly states (different protein features) and (iii) the relative contribution of a protein to a protein complex signal (protein complex features). It is important to keep in mind that all levels are quantified in relative terms and that the assessment will become increasingly powerful in future studies when comparing multiple biological conditions16.

One critical consideration is the level of redundancy in the prior protein connectivity information used as input for complex-centric analysis. While redundancy (e.g., in the form of larger and smaller protein sub-complexes) can significantly boost sensitivity and protein complex recovery, it also bears the potential to recover the same protein complex signal multiple times from the perspective of different, partially redundant queries. While the signal collapsing strategy in CCprofiler in principle removes such redundant complex assignments, FDR control does not propagate throughout this step of the workflow, and the results should be treated with more caution.

Finally, the protocol has the same caveats as most large-scale data analysis approaches. Since strict FDR control is necessary to warrant overall high-quality results, weaker, yet interesting signals might be missed. Therefore, it might still be advisable to manually inspect chromatograms of specific candidate proteins and protein complexes of interest. In case you can clearly determine the signals manually, an adjustment of the selected parameters for the CCprofiler analysis might be necessary to obtain optimal results. The obtained results can subsequently be leveraged to select protein complexes for further down-stream validation and characterization (e.g., by orthogonally validating their presence or investigating the structure–function relationship of unexpected complex variants in the cell).
Materials

Reagents

- HEK293 cell line (American Type Culture Collection, cat. no. CRL-1573). Preliminary work has shown that the protocol is also applicable to other cell lines\(^{18}\) and potentially also to tissue types (ongoing, unpublished work) ★ CRITICAL. The cell lines used in your research should be regularly checked to ensure that they are authentic and are not infected with mycoplasma.
- Cell culture dishes, 15 cm (Corning; Sigma-Aldrich, cat. no. CLS430599-60EA)
- Bicinchoninic acid (BCA) protein assay kit (Pierce; Fisher Scientific, cat. no. 23225)
- Injection vials, 32 × 11 mm (BGB Analytik, cat. no. PPSV0903K&090304)
- 96-well DeepWell plates (Nunc; Fisher Scientific, cat. no. 260251)
- 96-well MacroSpin plates C18 (Harvard Apparatus, cat. no. 74-5617)
- 96-well plate adhesive aluminum seals (VWR, cat. no. 60941-112)
- DMEM Thermo Fisher (Gibco; Fisher Scientific, cat. no. 670116)
- Penicillin-streptomycin-glutamine 100× (Gibco; Fisher Scientific, cat. no. 10378016)
- Fetal bovine serum (FBS) (Gibco; Fisher Scientific, cat. no. 26140079)
- Trypsin-EDTA 1×
- Base and SEC buffer: HEPES (Gibco; Sigma-Aldrich, cat. no. 11344041) and sodium chloride (Sigma-Aldrich, cat. no. S7653)
- Lysis buffer (lysis): Sodium fluoride (Sigma-Aldrich, cat. no. S6776), NP-40 detergent (Nonidet P-40, Sigma-Aldrich IGEPAL-630, cat. no. 18896), Na\(_3\)VO\(_4\) (Sigma-Aldrich, cat. no. S6508), PMSF (Sigma-Aldrich, cat. no. 78830) and protease inhibitor cocktail (Sigma-Aldrich, cat. no. P8340)
- SEC standard proteins (Phenomenex, cat. no. AL0-3042): thyroglobulin tetramer (1,398 kDa), thyroglobulin dimer (699 kDa), ovalbumin (44 kDa) and myoglobin (17 kDa)
- Acetonitrile, gradient grade (Sigma-Aldrich, cat. no. T6508)
- Internal retention time (iRT) normalization kit (iRTkit, Biognosys AG, cat. no. Ki-3002-1) featuring the following peptide sequences (iRT in parentheses): LGGNEQVTR (−28.308), GAGSSEPVTGLDAK (0.227), VEATFGVDESNAK (13.1078), YILAGVENSK (22.3798), TPVISGDPYKEYR (28.9999), TPVITGAPYEYR (33.6311), DGLDAASYAPVR (43.2819), ADVTPADSEWSK (54.969), GTFIIDPGGVIR (71.3819), GTFIIDPAAVIR (86.7152) and LFLQFGAQGSPFLK (98.0897)
- μl-volume UV-visible spectrophotometer (e.g., NanoDrop ND-1000, Thermo Fisher Scientific)
- LC-MS/MS system of nano-LC and DIA/SWATH-MS–enabled mass spectrometer (Ekisgent AS-2/1Dplus and AB SCIEX TripleTOF 5600+)
- PicoFrit self-pack columns and emitters (New Objective, cat. no. PF360-75-10-CE-5)
• Magic C18 Aq resin (3 μm, 200 Å; Michrom H254)
• Windows computer for file conversion with ProteoWizard software (http://proteowizard.sourceforge.net/)
• Workstation or server computer (any operating system) with ~300 GB of disk space, 12 GB random-access memory (RAM) and ≥6 threads/CPU cores. Please note that the number of threads is influencing processing speed only due to parallelized functions in both OpenSWATH and CCprofi ler.

Software to install on the Workstation or server computer
• Docker (e.g., https://docs.docker.com/docker-for-windows/)
• OpenSWATH Docker pipeline (see http://openswash.org/en/latest/docs/docker.html)
• R (≥v3.60, https://cran.r-project.org/bin/windows/base/) with packages devtools, data.table, ggplot2 and CCprofi ler (see ‘Equipment setup’)

Reagent setup
SEC buffer
The SEC mobile phase is 50 mM HEPES pH 7.5, 150 mM NaCl. 10× stock solutions can be stored at 4 °C for ≤8 weeks. Prepare 1,000 ml of SEC buffer per experiment to accommodate for system and column equilibration. Use milli-Q water. Before use, remove particles by 0.22-μm filtration and store at 4 °C. It is stable for ≤4 weeks.

HNN buffer
HNN buffer is equivalent to the SEC mobile phase, supplemented with 50 mM NaF for phosphatase inhibition. After 0.22-μm filtration, HNN buffer is stable at 4 °C for 8 weeks.

HNN lysis buffer
HNN lysis buffer is 0.5% (vol/vol) NP40, 50 mM HEPES (pH 7.5), 150 mM NaCl, 50 mM NaF, 200 μM Na3VO4, 1 mM PMSF and 1× protease inhibitor cocktail. The buffer is to be prepared fresh for each experiment by dilution of stock solutions into HNN buffer. Aliquots of stock solutions are prepared and stored as follows. 20% (vol/vol) NP-40 in milli-Q water can be stored at room temperature (RT, 23–25 °C) in the dark (wrap with aluminum foil) for several weeks. 200 mM Na3VO4 in H2O (100×) is divided into aliquots and stored at −20 °C. 1 M PMSF in 70% (vol/vol) ethanol (EtOH) (100×) is divided into aliquots and stored at −20 °C. Sigma protease inhibitors are divided into aliquots of 20 μl and stored at −20 °C. Aliquots are stable for at least 3 months.

C18 elution buffer
C18 elution buffer is 50% (vol/vol) acetonitrile (ACN) in 0.1% (vol/vol) formic acid (FA). This can be stored at RT for ≤3 months.

Sample resuspension solution
Sample resuspension solution is 2% (vol/vol) ACN in 0.1% (vol/vol) FA with Biognosys iRT peptides spiked in at a ratio of 1:20. This can be stored at RT for ≤3 months.

NanoLC mobile phase A
NanoLC pump mobile phase A is 2% (vol/vol) ACN in 0.1% (vol/vol) FA. Mobile phase should be freshly prepared.

NanoLC mobile phase B
NanoLC pump mobile phase B is 90% (vol/vol) ACN in 0.1% (vol/vol) FA. Mobile phase should be freshly prepared.

Equipment setup
Off-line complex fractionation by SEC
Note that SEC reproducibility can be compromised by differences in flow due to leakage or increased backpressure from clogged guard column cartridges. Monitor typical overall system backpressure and in-run backpressure to spot and solve leaks. We recommend employing two guard cartridges in line and replacing the upstream cartridge as soon as the pressure difference (∆p) across the guard column exceeds 10 bar at 500 μl/min. To avoid damage to the 3-μm bead SEC column, avoid sudden pressure changes by adjusting flow rates only in small increments of 100 μl/min and allowing...
~5 s for pressures to adjust. The column is equilibrated by 10 column volumes (150 ml) of SEC buffer. In proteome-wide SEC fractionations, secondary interactions with the stationary phase of a subset of analytes and consequent column conditioning and washout effects cannot be avoided. Therefore, we recommend pre-conditioning the column with a lysate similar to the lysate to be analyzed in SEC-SWATH-MS to ensure consistent analyte elution volume and recovery in the fractionation. Mild lysates concentrated for SEC column conditioning in 1,000-μg aliquots can be stored at −80 °C for several months. Thaw an aliquot, spin out precipitates by 5 min of centrifugation of 16,900g (4 °C) and run in SEC. After column conditioning (~90 min to allow full baseline equilibration of the OD signal), analyze the aqueous SEC standard sample to finish setup for analysis of the real sample(s).

Fractionations are run at 500 μl/min to minimize shear forces and with column temperature controlled at ≤4 °C to minimize dilution-induced complex disassembly. Mobile phases are stored at RT and cooled by flow through the temperature-controlled autosampler module as well as heat exchanger units before entering the guard and analytical columns. To accommodate both the pre-column and the analytical column, we modified the housing of the column compartment with a drill. Alternatively, if no modifications are possible or if larger-capacity SEC columns are to be employed, temperature can be controlled by submerging both the pre-column and the main column in an ice-water bath.

**On-line nanoLC-MS/MS**

Peptide samples are loaded onto a self-packed C18 reversed-phase column (75-μm i.d. PicoFrit emitter packed with 20-cm Magic AQ 3-μm C18 resin) at 300 nl/min and subsequently eluted by a linear gradient of 3–35% mobile phase B in mobile phase A over 120 min at 300 nl/min, with direct electrospray into the ion source and mass spectrometer. Other comparable C18 phases and nanoLC setups can be employed. The TripleTOF 5600+ mass spectrometer is operated in either DDA mode (also termed information-dependent acquisition mode) or DIA/SWATH2.0 acquisition mode. For detailed acquisition parameters, see Tables 1, 2 and 3. Other mass spectrometric platforms, such as the TripleTOF 6600, have been used successfully to record high-quality SEC-SWATH-MS datasets.

**Windows PC for file conversion**

For file conversion, a Windows computer with a recent version of the ProteoWizard suite (≥ version: 3.0.19228-a2fc6eda4) is required. Download and Install ProteoWizard from http://proteowizard.sourceforge.net/ as described55.

**Workstation computer for peptide-centric SWATH-MS analysis**

In this protocol, we employ a docker container that provides a stable solution for running peptide-centric scoring by OpenSWATH, PyProphet and TRIC on different computing systems. The workflow presented here has been tested on Linux, Windows and OSX environments. First, it is necessary to install Docker (https://docs.docker.com/). Note that it might, depending on the dataset and library size, be necessary to extend the resources allocated to the docker software. We recommend minimally allocating six CPUs, ~12 GB memory and ~12 GB of disk image size. If a task running within a docker container is suddenly ‘killed’ without a more specific error, this usually means that not enough memory was allocated. Settings can be changed when clicking on the docker symbol in the taskbar, selecting the settings option and going to the advanced tab. To test the successful docker installation, open a command line interpreter on your computer and type the following command:

```
docker run hello-world
```

Now install the OpenSWATH docker container:

```
docker pull openswath/openswath:0.1.2
docker run -u 0 -dit --name openswath -v $PWD/:/data openswath/openswath:0.1.2
```

Test if the OpenSWATH docker installation worked:

```
docker exec openswath echo hi there, openswath container is happy and alive
```
R environment and CCprofiler installation

All data analysis in R can be performed either on a local computer or on a cluster system. To install R, download the latest release version of R (≥3.6.0) from http://cran.r-project.org/ and install it according to the R installation and administration manual (https://cran.r-project.org/doc/manuals/R-admin.html). To install the CCprofiler package from GitHub, you need to start the R program. For both Windows and OSX, this means double-clicking on the R application icon. On UNIX-like systems, you need to type ‘R’ in a shell prompt. Users may also want to consider using the RStudio environment (https://rstudio.com/).

Once in the R environment, run the following commands to first install and then load the devtools, data.table and CCprofiler packages:

```r
install.packages('devtools')
library('devtools')
install.packages('data.table')
library('data.table')
install_github('CCprofiler/CCprofiler')
library('CCprofiler')
```

Procedure

Module 1: isolation of native proteome, SEC fractionation and preparation for MS analysis

Cell culture and harvest ★ Timing ~6 d

1. Culture cells as applicable to the respective cell type. If using HEK293 cells, culture cells in DMEM medium supplemented with 10% (vol/vol) FBS and 50 μg/ml (wt/vol) penicillin-streptomycin in 15-cm cell culture dishes, incubating at 37 °C, 5% CO₂. To establish a log-linearly growing cell population, split the cells twice at a ratio of 1:2 using 1x trypsin-EDTA for 5 min at 37 °C.

2. Harvest the cells at ~80% confluency, as determined by visual inspection under the microscope. Harvest enough cells to enable you to extract ≥1 mg of total protein for every sample that you want to analyze (in the case of the HEK293 cell line, this is 7 × 10⁷ cells). Harvest cells on ice in ice-cold PBS buffer containing 5 nM EDTA using pipette flow (sufficient in the case of HEK293 cells) or a plate scraper into a 15-ml Falcon tube. Spin at 4 °C and 500 g for 5 min at 37 °C.

Native lysis and fractionation by size exclusion chromatography ★ Timing 1 d

3. To monitor SEC stability and to calibrate the apparent MW per SEC fraction, analyze 5 μl of the SEC column performance check standard before and after the SEC experiment.

4. Lyse cells sufficient to extract ≥1 mg of total protein (in the case of the HEK293 cell line, 7 × 10⁷ cells). Lyse cell pellets snap-frozen in Step 2 by freeze-thawing into 1 ml of HNN lysis buffer. Thaw and dissolve the frozen pellet by pipetting up and down 20 times. Incubate on ice for 5 min. Other cell or potentially also tissue types may be used, whereas input amounts need to be adapted based on cell size or protein yield with a minimal pure protein amount of 600 μg required as input to SEC fractionation, with concentration determined colorimetrically (e.g., using the Pierce BCA protein assay kit). This corresponds to ~2 mg when protein concentration is estimated by OD₂₈₀ measurements, which are confounded by other molecules in the sample but used here for the sake of processing speed.

★ CRITICAL STEP ★ To adapt the protocol for other sample types, alternative lysis procedures may be required to access the respective proteomes in assembled form. To avoid thermal disruption of protein complexes, we recommend working in the cold as much as possible. While freeze-thawing will likely suffice for most cell lines, more robust samples may require stronger mechanical disruption procedures, such as dounce-homogenization, bead-beating under temperature control, use of mortar and pestil under liquid nitrogen or, if available, freeze milling. Such sample types could be tissue specimens from any organism or more sturdy single-cell organisms such as yeast or prokaryotes. However, these considerations extend beyond the protocol presented here.
Fill the lysate to a volume of 2 ml with HNN lysis buffer and distribute to two ultracentrifuge tubes. Balance weight on a fine balance with HNN lysis buffer.

Transfer to the pre-cooled centrifuge rotor and clarify by 15 min of ultracentrifugation (100,000g, 4 °C, 55,000 rpm on TLA120.2 rotor).

Pre-cool two Amicon Ultra-4 centrifugal filter units on ice. Transfer 300 µl of the cleared lysate to each Amicon device and exchange buffer to HNN buffer as follows.

**CRITICAL STEP** Avoid transfer of lipids from the top layer of the supernatant by aspirating the cleared lysate from 1 cm below the liquid surface.

Exchange the buffer to HNN (50 mM HEPES (pH 7.5), 150 mM NaCl, 50 mM NaF) at a final ratio of 1:50 in three dilution and re-concentration steps to avoid large dilution steps in the interest of complex integrity. All centrifugations for this action are performed at 3,220g and 4 °C.

- Centrifuge for 5 min (final volume above filter: ~200 µl).
- Dilute 1:5 in HNN (add 800 µl), flush membrane and centrifuge for 10 min (final volume: ~250 µl).
- Dilute 1:5 in HNN (add 1,000 µl), flush membrane and centrifuge for 10 min (final volume: ~250 µl).
- Dilute 1:2 in HNN (add 250 µl) and flush the membrane.
- Centrifuge for 5 min (volume: ~150 µl) and flush the membrane with the sample above the membrane.
- Centrifuge for 5 min. The final volume per tube: ~50–80 µl.

Remove precipitates by centrifugation at 16,900g and 4 °C for 5 min, transferring the supernatant, leaving 10 µl to waste to avoid aspiration of any precipitates, to a pre-cooled injection vial.

**CRITICAL STEP** Local precipitation occurs at and blocks the filtration membrane. It is therefore important to flush the membrane with the dilution buffer using a 200-µl pipette tip to achieve thorough rinsing of the membrane.

Measure the concentration of the lysate by UV-visible photospectrometry (Nanodrop spectrophotometer) against a reference sample of HNN lysis buffer in HNN buffer (1:50), approximating 1 OD280 = 1 µg/µl protein concentration. The measured concentration should typically be between 20 and 30 µg/µl.

**CRITICAL STEP** The concentration read by UV-visible photospectrometry is confounded by other compounds with absorbance at 280 nm. Based on colorimetric methods (BCA assay), the protein loading is ~three- to-fourfold lower than approximated by UV-visible photospectrometry (Fig. 2b; i.e., the real protein concentration will be between 5 and 10 µg/µl). Despite this, we choose to use fast UV-visible reading to align sample-loading amounts as it enables faster sample preparation for fractionation with minimized impact on complex integrity, which is preferable over BCA or similar quantitative assays with significant incubation times.

Subject 1,000 µg of protein (corresponding to 33–50 µl of the concentrated lysate) to SEC fractionation at 500 µl/min. Ensure that the chromatographic system and column show reproducible and expectable performance in the fractionation of the SEC column performance check standard before and after the analysis. Collect fractions in the expectable elution range from 10 to 28 min at 0.19 min per fraction into a cooled 1 ml 96-well deep-well plate.

Repeat Step 10 and collect the second batch of fractions in a new 96-well plate.

**CRITICAL STEP** The reason to split the separation into two runs is to obtain sufficient material for DDA-MS and SWATH-MS analysis, a limitation that may be overcome by optimizing the sensitivity of protein digestion and peptide yield, using larger-capacity SEC columns or using more-sensitive MS instrumentation. The upper limit of 1,000 µg to be separated on the 300 × 7.8 mm Yarra columns was determined by reducing injection amounts such that the peak of residual NP-40 detergent (visible at OD280 at the end of the fractionation; compare Fig. 2b) is effectively separated from the protein separation range of interest. In other words, injecting >1,000 µg of the soluble complex preparation led to NP-40 eluting with the proteins of interest and interference with MS analysis. Therefore, the total required amount of protein was fractionated in two consecutive fractionations.

Check the UV-visible profiles of the two SEC runs of the same lysate, and if they are the same, pool the collected fractions across the two replicate injections to obtain one set of fractions.

**CRITICAL STEP** It is important to also sample chromatographic fractions of the void volume peak, even if the information of contained analyte size is reduced. This is especially important for quality control measures of the overall global proteome assembly state of the investigated cell system (observed total MS signal in assembled versus monomeric SEC range). In addition, the peak detection algorithms employed in downstream protein and protein complex detection benefit from
complete elution profiles, including shoulder regions of detectable peaks. The right boundary of the relevant protein elution range can be established empirically by SDS-PAGE analysis of the late fractions (>fraction 70). We suggest using the elution volume of the small molecule uridine contained in the SEC standard sample. We recommend sampling right up until the uridine peak has been eluted and sampled, because a subset of proteins and complexes may display secondary interactions with the stationary phase and thus delayed elution in this fraction range.

13 Transfer an aliquot of the unfraccionated sample to the collection plate. Pipette 1/40th of the volume injected for SEC (25 µg by OD280) into free wells (H11 and H12) and fill to 200 µl with SEC buffer to align digest conditions with the individual SEC fractions. ▲ CRITICAL STEP The data acquired from the unfraccionated mild proteome extract is used in the PyProphet machine-learning step in peptide-centric analysis. It is used to generate a single scoring function that is applied across all chromatographic fractions to ensure aligned scoring and consistent quantification of peptides across all chromatographic fractions. It is important to include the aliquot of the unfraccionated sample with the other samples undergoing proteomic analysis to ensure comparable digest conditions as for the chromatographic fractions.

▲ PAUSE POINT Undigested SEC fractions can be stored at −80 °C for several weeks. Optionally, if extended storage is desired, it is recommended to denature proteins by boiling in sodium deoxycholate (next step) before freezing for storage.

Module 2: data acquisition by bottom-up proteomics analysis of all collected fractions

Tryptic digest and C18 cleanup of chromatographic fractions for MS analysis ● Timing 12 h + 4 h

14 Denature proteins by adding sodium deoxycholate to 1% (vol/vol) (20 µl from 10% (vol/vol) stock solution in LC grade H2O) and incubate 5 min in a hot water bath (95 °C).
▲ CRITICAL STEP Ensure that the plate is properly sealed before incubation in the water bath to avoid sample loss, contamination or excessive evaporation.

15 Let the plate cool to RT and centrifuge at 500g for 1 min at RT to collect the liquid at the base of the plate.

16 Reduce proteins by adding TCEP to 5 mM (22 µl from 50 mM solution, 1:10 dilution of 500 mM stock in ammonium bicarbonate 50 mM, pH 8.8). Incubate 30 min at RT.
▲ CRITICAL STEP Ensure that the TCEP stock solution is titrated to pH 8.8 to avoid acidification of the samples and premature precipitation of sodium deoxycholate.

17 Alkylate proteins by adding iodo-acetamide to 10 mM (24 µl from 100 mM stock). Incubate 20 min at RT in the dark.
▲ CRITICAL STEP Work in reduced light conditions and incubate in the dark due to iodo-acetamide light sensitivity.

▲ CRITICAL STEP Ensure that the pH is ≥8.0 to avoid gel formation or partial precipitation of deoxycholate during the digest. Test the samples for gel formation using a pipette tip, and if very high viscosity or formation of a gel are observed, adjust the pH by adding NaOH (in steps of 5 µl of 2 M stock solution until the samples display low viscosity and pH 8.0–8.5).

18 Add 0.2 µg trypsin (Promega) per fraction (2 µl of 0.1-µg/µl stock in trypsin buffer). Re-seal the plate, shake, spin down for 1 min at RT and 2,000g and incubate overnight at 37 °C.

19 Stop the digest and precipitate deoxycholic acid by adding TFA to 1% (vol/vol), ACN to 1% (vol/vol) (26 µl of 10% TFA/10% ACN (vol/vol) in LC grade H2O stock solution). Close and mix the plate thoroughly using a new plate seal and 10 inversions. Spin down for 1 min at 2,000g.

20 Prepare a MacroSpin plate for C18 cleanup. Tap the plate to loosen resin material and spin down for 1 min at 1,000g. Activate resin by adding 200 µl of ACN per well and centrifuging at 1,000g for 1 min. Equilibrate the resin by three washes with 150 µl of 5% (vol/vol) ACN/0.1% (vol/vol) FA, spinning at 1,000g for 2 min. Discard the washing solution from the collection plate.

21 Directly before loading the samples for C18 cleanup, pellet the precipitated deoxycholic acid for 10 min at 3,220g. Transfer 80% (220 µl) of the cleared supernatant onto the equilibrated C18 resin.
▲ CRITICAL STEP Ensure minimal transfer of precipitate onto the C18 resin to avoid sample contamination.

22 Load samples at 1,000g for 2 min. To maximize recovery, re-load the flow-through onto the C18 resin a second time. Keep the flow-through for potential troubleshooting.

23 Wash the C18 resin by 3 × 200 µl of 5% (vol/vol) ACN/0.1% (vol/vol) FA, spinning at 1,000g for 1 min each.
24 Elute the samples into a fresh collection plate with 2 × 150 µl of 50% (vol/vol) ACN/0.1% (vol/vol) FA.
25 Dry samples in a SpeedVac equipped with a plate rotor and adequate tara plates filled with the same volume of C18 elution buffer (45 °C, 0.2 atm, ~4 h).

**Pause Point** Dried peptide samples can be stored for several weeks at −20 or −80 °C.

**MS analysis**

- **Timing** 12 h (quality control) + 14 d (DIA only) or 28 d (DIA + DDA)

26 Re-suspend dried peptide samples in 18 µl of 2% (vol/vol) ACN/0.1% (vol/vol) FA, supplemented with iRT calibration peptides (iRT kit, Biognosys; 1:20 dilution as opposed to manufacturer’s instruction of 1:10 to accommodate larger injection volumes). The spiked-in iRT peptides allow the normalization of retention times across different LC-MS/MS runs and enable the streamlined generation of spectral libraries and queries of peptides from repository-scale spectral libraries in the DIA/SWATH-MS data maps. Re-suspend the samples by 5-min sonication in an ice-cooled water bath to avoid sample heating and evaporation.

27 Collect the liquid and remove potential residual deoxycholate by centrifugation at 3,220g for 5 min. Transfer 16 µl of the sample to MS injection vials.

**Critical Step** Transfer the peptide samples by pipetting at an angle and leaving ~2 µl to avoid transfer of potential residual deoxycholate precipitate from the lowest points of the wells.

28 Before analyzing the full set of fractions, test the sample set quality by analyzing 2 µl of the unfractionated sample and the two fractions with the highest absorbance at OD280 as monitored during SEC fractionation (in our chromatographic setup, fractions 5 and 50).

Judge sample quality based on the following criteria: (i) no increase in chromatographic backpressure, (ii) TIC signal intensity in SWATH64vw mode is ≥2 × 10^7 (120-min gradient) (Fig. 2c) and (iii) the m/z map is well populated with isotopic envelopes. To acquire the full dataset, maximize sample injection volumes aiming for a TIC signal intensity of 1 × 10^8 in the highest-abundant SEC fraction (in the HEK293 case, fraction 50 and corresponding to an injection volume of 4 µl).

**Troubleshooting**

29 If a project-specific spectral library should be generated, analyze each fraction in both DIA/SWATH and DDA mode.

**Critical Step** While datasets acquired exclusively in SWATH acquisition mode can typically be interpreted using spectral libraries from public repositories, it is important that the library employed for interpretation be representative of the cell or tissue type that is being analyzed. Depending on the availability of such libraries and the research question at hand, it might further be preferable to generate a project-specific spectral library by DDA of a subset of or the full sample set analyzed by SEC-SWATH-MS.

**Peptide-centric SWATH-MS analysis**

- **Timing** 3 d

**Critical** For the data analysis, we employ a docker container (see Equipment setup) that provides a stable solution for running peptide-centric scoring by OpenSWATH, PyProphet and TRIC on any computing system. Example files and a script including all processing steps are provided in our GitHub repository (https://github.com/CCprofiler/SECSWATH_PeptideCentricAnalysis.git). If you wish to perform the computational, peptide-centric analysis, we recommend that you modify the scripts provided in the GitHub repository or copy the commands from there. Because the computer is very sensitive to correctly set white spaces and line breaks, copying code from an online document or a .pdf file might lead to unexpected errors, especially for an inexperienced user.

**Prepare data for peptide-centric analysis**

30 Create a data analysis folder: open a command line interpreter and clone and enter our analysis folder template from GitHub:

```bash
git clone https://github.com/CCprofiler/SECSWATH_PeptideCentricAnalysis.git
```

```bash
cd SECSWATH_PeptideCentricAnalysis
```

31 The first stage in preparing all the required input data for peptide-centric analysis is MS file conversion and centroiding. On the conversion computer, use MConvert to convert and centroid .wiff raw files into .mzML or mzXML format. Open MConvertGUI. Under ‘Files/browse’, select
the .wiff files. Under ‘Options’, leave the defaults and activate in addition ‘Package in gzip’. Under ‘Filters’, select ‘Peak Picking’. Under ‘Algorithm’, select ‘Vendor’. Under ‘MS Levels’, enter ‘1-2’. Once you have made the appropriate selections, hit ‘Add’. Start the conversion (button in the lower right).

**Critical Step** The centroiding significantly reduces file size and processing time and is highly recommended, in particular, if peptide-centric analysis is to be performed on a personal or laptop computer.

32 Once the conversion is finished, move the .gz file(s) to the peptide-centric analysis computer and into the folder

```
SECSWATH_PeptideCentricAnalysis/data_dia/
```

33 Then, move the .gz files generated from the unfractionated sample into the subfolder

```
SECSWATH_PeptideCentricAnalysis/data_dia/unfractionated_secinput/
```

34 Prepare a spectral library with information about the retention times of the calibration peptides (iRT spike-in or common iRT (ciRT) peptide set). Example iRT and ciRT libraries are provided in the data_library folder in the cloned GitHub repository.

35 Prepare a file specifying the SWATH window settings. An example file with SWATH window settings is provided in the data_library folder in the cloned GitHub repository (Table 3).

36 Prepare a spectral library. Either create a sample-specific spectral library according to the previously published protocol by Schubert et al.29 or download a public library such as the CAL that we used for our analysis here34. To download the CAL, type the following in the command line interpreter:

```
wget -O data_library/spectrast2tsv.tsv https://db.systemsbiology.net/sbeams/cgi/downloadFile.cgi?name=phl004_canonical_s64_osw.csv;format=tsv;tmp_file=8becf7ae782dd305c0eade59f282bcd1;raw_download=1
```

37 Initialize the OpenSWATH docker container (see Equipment setup or follow instructions in https://github.com/CCpro/SECSWATH_PeptideCentricAnalysis/blob/master/SECSWATH_PeptideCentricAnalysis.sh)

```
docker attach openswath
```

38 First, convert the selected spectral library to .pqp file format, which is recommended for OpenSWATH.

```
TargetedFileConverter -in /data/data_library/spectrast2tsv.tsv -out /data/data_library/spectrast2tsv.pqp
```

39 Then, generate decoys for scoring and FDR estimation in PyProphet

```
OpenSwathDecoyGenerator -in /data/data_library/spectrast2tsv.pqp -out /data/data_library/spectrast2tsv_td.pqp
```

**Peptide-centric signal detection with OpenSWATH**

40 First, run OpenSWATH on the unfractionated input sample(s).

```
for file in /data/data_dia/unfractionated_secinput/*ML.gz; do 
 bname=$(echo ${file##*/} | cut -f 1 -d '.'); 
 OpenSwathWorkflow 
   -in /data/data_dia/$bname.*ML.gz 
   -tr /data/data_library/spectrast2tsv_td.pqp 
   -tr_irt /data/data_library/irtkit.TraML 
   -min_upper_edge_dist 1 
```

NATURE PROTOCOLS | VOL 15 | AUGUST 2020 | 2341-2386 | www.nature.com/nprot
41 Then, run OpenSWATH on the fractionated samples.

```bash
for file in /data/data_dia/*ML.gz; do 
  bname=$(echo ${file##*/} | cut -f 1 -d '.'); 
  OpenSwathWorkflow 
  -in /data/data_dia/$bname.*ML.gz 
  -tr /data/data_library/spectrast2tsv_td.pqp 
  -tr_irt /data/data_library/irtkit.TraML 
  -min_upper_edge_dist 1 
  -batchSize 1000 
  -out_osw /data/results/$bname.osw 
  -rt_extraction_window 600 
  -mz_extraction_window 30 
  -ppm 
  -threads 6 
  -use_ms1_traces 
  -Scoring:Scores:use_ms1_mi 
  -Scoring:Scores:use_ms1_score; done
```

**CRITICAL STEP** OpenSWATH creates several warnings and errors that can be ignored when analyzing SEC-SWATH-MS datasets, including: (i) warning 'windows were sparse' and/or 'empty chromatogram' (sparsity of certain windows is expected for some fractions, especially in the beginning and end of the SEC) and (ii) error 'Transition does not have a corresponding chromatogram'.

**Peptide-centric scoring with PyProphet and TRIC**

42 First, train a scoring model by running a PyProphet analysis of the unfractionated input sample.

```bash
pyprophet score --threads 6 --in=/data/results/unfractionated_secinput/unfractionated_secinput.osw --out=/data/results/unfractionated_secinput/model.osw --level=ms1ms2
```

43 Then, apply the model to score peak groups in all fractionated samples accordingly.

```bash
for file in /data/results/*.osw; do 
  bname=$(echo ${file##*/} | cut -f 1 -d '.'); 
  pyprophet score --in=/data/results/$bname.osw --apply_weights=/data/results/unfractionated_secinput/model.osw --level=ms1ms2; done
```

Exporting of output files:

```bash
for file in /data/results/*.osw; do 
  bname=$(echo ${file##*/} | cut -f 1 -d '.'); 
```
pyprophet export --in=/data/results/$bname.osw \
--out=/data/results/$bname.tsv \
--max_rs_peakgroup_qvalue=0.1 \
--no-transition_quantification \
--format=legacy_merged; done

**▲ CRITICAL STEP** We advise manually checking if .tsv output files are actually written for all runs.

Plotting of all score distributions:

```
for file in /data/results/*.osw; do \
  bname=$(echo ${file##*/} | cut -f 1 -d '.'); \
  pyprophet export --in=/data/results/$bname.osw \
  --format=score_plots; done 
```

44 Perform TRIC-based feature alignment across all SEC fractions.

```
feature_alignment.py \ 
--in /data/results/*.tsv \ 
--out /data/results/feature_alignment.tsv \ 
--out_matrix /data/results/feature_alignment_matrix.tsv \ 
--method LocalMST \ 
--realign_method lowess \ 
--max_rt_diff 60 \ 
--mst:useRTCorrection True \ 
--mst:Stdev_multiplier 3.0 \ 
--target_fdr -1 \ 
--fdr_cutoff 0.05 \ 
--max_fdr_quality 0.1 \ 
--alignment_score 0.05
```

**Module 3: SEC-SWATH-MS data processing and complex-centric analysis in CCprofiler**

**● Timing 2 d**

**▲ CRITICAL** Module 3 of the procedure describes how to use the open-source CCprofiler R-package to extract information about the global proteome assembly state and specific protein complexes from co-fractionation MS experiments, here generated by SEC-SWATH-MS. The analysis includes: data preparation (Steps 45–49), loading and inspection of the input data in R (Steps 50–56), creation of a CCprofiler traces object (Steps 57–60), peptide traces pre-processing (Steps 61–64), peptide traces inspection (Steps 65 and 66), protein quantification (Steps 67–69), overall quality control (Step 70), automated parameter selection (Steps 71–73), protein-centric analysis (Steps 74–78) and complex-centric analysis (Steps 79–89). All CCprofiler analysis steps are also provided as an R-script that performs the presented analysis based on the exemplary HEK293 SEC-SWATH-MS dataset. The R-script can easily be adapted to other datasets by changing the input files (Steps 34 and 35). All exemplary data and the script are available on GitHub: https://github.com/CCprofiler/SECSWATH_ComplexCentricAnalysis (Supplementary CCprofiler Manual). If you wish to perform the CCprofiler analysis, we recommend that you modify the scripts provided in the GitHub repository or copy the commands from there. Similar to the previous section, the reason for this is that the computer is very sensitive to correctly set white spaces and line breaks. Copying code from an online document or a .pdf file might therefore lead to unexpected errors, especially for an inexperienced user.

To set up your work environment, you can clone the GitHub repository by:

```
git clone https://github.com/CCprofiler/SECSWATH_ComplexCentricAnalysis.git 
```

cd SECSWATH_ComplexCentricAnalysis
Prepare data for CCprofiler import

**CRITICAL** Prepare all necessary data that need to be loaded into R for the CCprofiler analysis. A summary of all required input data is provided in Table 4. For convenience, we recommend saving all input data in the same directory where you want to perform the analysis. All data necessary and used for this protocol are provided in the GitHub repository and will be available in the SECSWATH_ComplexCentricAnalysis folder after you clone it (see above).

45 Prepare quantitative peptide-level data derived from either OpenSWATH (A) or another source (B).

(A) **Quantitative peptide matrix generated by OpenSWATH (as described in Part 2)**

1. Import the output table from TRIC directly into CCprofiler (see ‘feature_alignment.tsv’ or ‘quantData_OpenSWATH.rds’ (already in R data format)).

(B) **Quantitative peptide matrix generated by any software tool**

1. Remove decoys.
   - **CAUTION** Decoys might be valuable for certain processing steps downstream (e.g., selecting a SPC-based FDR cutoff). We have specifically tested the propagation of decoys for datasets processed by an OpenSWATH-based workflow. If other data processing tools have been used, the decoys should be treated with caution. To be on the conservative side, we would generally recommend removing the decoys.

2. Remove non-proteotypic peptides.

3. Bring data in either long or wide format. The required column names for long format are protein_id, peptide_id, filename and intensity (see ‘examplePCPdataLong.tsv’). The required column names for wide format are protein_id, peptide_id, <filename1>, <filename2>, …, <filenameN> (see ‘examplePCPdataWide.tsv’).

46 In addition to the quantitative peptide matrices, CCprofiler requires a fraction annotation table that maps each filename to a given chromatographic fraction number. Create this table using the following required column names: filename and fraction_number (see ‘exampleFractionAnnotation.tsv’).

- **CAUTION** The filenames used in the fraction annotation table need to match the filenames in the quantitative matrix exactly. Furthermore, the fraction_number entries need to start with 1 and continuously increase in integer steps of 1 until the last sampled fraction.

47 (Optional) For native complex separation via SEC, generate an MW calibration table by measuring the apex fractions of an external standard set of reference proteins with known MWs fractionated on the same SEC setup. By providing such an MW calibration table, CCprofiler can establish a transformation function based on the log-linear relationship between elution fractions and apparent MWs inherent to SEC, thus enabling the annotation of all sampled fractions with an apparent MW. The required column names in the calibration table are std_weights_kDa and std_elu_fractions (see ‘exampleCalibrationTable.tsv’).

48 Select a trace annotation table to provide additional information about peptides and proteins, e.g., adding the gene names and monomeric MW from UniProt (https://www.uniprot.org/) (see ‘exampleTraceAnnotation.tsv’). Adding the information on monomeric MWs of the analyzed proteins is critical for the assignment of proteins to monomeric or complex-assembled state from SEC datasets with calibrated apparent MW and is required for the assessment of global proteome assembly states.

- **CAUTION** The protein_id column in the quantitative matrix needs to match one of the column entries in the annotation table. Typically, the common entry is the UniProt identifier.

49 Select prior protein connectivity information. This can be provided either in the form of defined protein complexes (e.g., as annotated in CORUM[44,46]) or binary interaction networks generated by various approaches, as, for example, the BioPlex[2,3] or StringDB[45,46] networks. This component is necessary for downstream detection of protein complexes by complex-centric analysis (Steps 79–89). For defined complex hypotheses, a table with defined complexes should contain
the following columns: complex_id, complex_name and protein_id (see 'corumComplexHypotheses Redundant.csv'). For binary protein-protein interaction networks, the format is a table with two columns: a and b. Both columns contain protein identifiers, and each row represents a binary connection (an 'edge') in the interaction network (see 'BioPlexPPIs.tsv' on GitHub: https://github.com/CCprofiler/SECSWATH_ComplexCentricAnalysis/blob/master/BioPlexPPIs.tsv).

! CAUTION The protein_id/a & b entries need to correspond to the protein_id in the quantitative matrix (e.g., UniProt identifiers).

Load input tables into R and inspect

50 Load libraries in R.

library(data.table)
library('CCprofiler')

51 Set working directory to the location where all files are stored and define that PDF plots should be written.

setwd("SECSWATH_ComplexCentricAnalysis")
plotPDF <- TRUE

52 Load and inspect the quantitative peptide data. The peptide-level MS data may have been derived from OpenSWATH or from another software tool. For alternative software tools, the data can be prepared to correspond to either a long or a wide format (Step 45B). CCprofiler will automatically detect if peptide tables are in long or wide format. The syntax for importing data of the different formats is shown below.

| OpenSWATH            | quantData_OpenSWATH <- readRDS("quantData_OpenSWATH.rds") |
|----------------------|---------------------------------------------------------------|
| Other software,      | quantData_long <- fread("examplePCPdataLong.tsv")           |
| long format          | head(quantData_long)                                         |
| Other software,      | quantData_wide <- fread("examplePCPdataWide.tsv")           |
| wide format          | head(quantData_wide[,1:5])                                   |

53 Load and inspect the fraction annotation table.

fractionAnnotation <- fread("exampleFractionAnnotation.tsv")
head(fractionAnnotation)

54 Load and inspect the calibration table.

calibrationTable <- fread("exampleCalibrationTable.tsv")
calibrationTable

55 Load and inspect the trace annotation table.

uniprotAnnotation <- fread("exampleTraceAnnotation.tsv")
head(uniprotAnnotation)

56 Load and inspect protein connectivity information. To use the defined complex hypotheses from the Corumdatabase:

corumComplexes <- fread("corumComplexHypothesesRedundant.csv")
head(corumComplexes)

To use the binary protein-protein interaction network from BioPlex (v1.01, http://bioplex.hms.harvard.edu):

BioPlexPPIs <- fread("BioPlexPPIs.tsv")
head(BioPlexPPIs)
Import peptide level data into CCprofi

ler traces format and annotate

57 Import the quantitative peptide matrix as a traces object. The traces object is the main data class used in the CCprofi

ler package. It stores the quantitative profiles ('traces') of peptide or protein intensities across the analyzed chromatographic fractions. In addition, a traces object can store specific information about each of the peptides, proteins and chromatographic fractions. As the analysis proceeds, more information will be added to the traces object.

```
OpenSWATH
pepTraces <- importFromOpenSWATH(data = quantData_OpenSWATH, annotation_table = fractionAnnotation, verbose = FALSE)
```

Other software, long format
```
pepTraces_exampleSubset_long <- importPCPdata(input_data = quantData_long, fraction_annotation = fractionAnnotation, rm_decoys = FALSE)
```

Other software, wide format
```
pepTraces_exampleSubset_wide <- importPCPdata(input_data = quantData_wide, fraction_annotation = fractionAnnotation, rm_decoys = FALSE)
```

58 Perform MW calibration based on a provided calibration_table (Fig. 3a):

```
calibration <- calibrateMW(calibration_table = calibrationTable, PDF = plotPDF).
```

59 Annotate traces with the apparent MW associated with each SEC fraction as extrapolated from the standard protein MWs and associated elution fraction numbers:

```
pepTraces <- annotateMolecularWeight(traces = pepTraces, calibration = calibration)
```

! CAUTION Apparent MW calibration is of limited accuracy as, inherent to the analytical procedure, analyte shape and propensity for unintended secondary interaction with the stationary phase affect elution volumes/fraction number and inferred apparent MW. Predictions, especially those outside the range of standard protein elution, should be interpreted with caution.

60 Annotate traces with information from UniProt.

```
pepTraces <- annotateTraces(traces = pepTraces, trace_annotation = uniprotAnnotation, traces_id_column = "protein_id", trace_annotation_id_column = "Entry")
```

Pre-process traces object to increase data quality

61 (Optional): Detect and impute missing values. In most proteomics pipelines, zero-intensity values indicate either that the signal is missing at random (no detection due to technical reasons such as interference from other peptides) or missing not at random (no detection due to cellular concentrations below the detection limit). We suggest that a zero value is likely missing at random in case a quantitative (non-zero) signal has been detected in both the previous and following fractions. The detected missing-at-random values are subsequently imputed by a spline fit across the fractionation dimension. This is a three-step process. Convert zeros in missing-at-random value locations to NA:

```
pepTracesMV <- findMissingValues(traces = pepTraces, bound_left = 2, bound_right = 2, consider_borders = FALSE).
```

Impute NA values by fitting a spline:

```
pepTracesImp <- imputeMissingVals(traces = pepTracesMV, method = "spline")
```

```
Plot imputation summary:

```r
plotImputationSummary(traces = pepTracesMV,
                      tracesImp = pepTracesImp,
                      max_n_traces = 5,
                      PDF = plotPDF).
```

Note that in the original complex-centric study of the HEK293 proteome, no missing values were imputed. Generally, quantitative matrices from SWATH-MS, particularly with TRIC alignment, display only a few missing values, and imputation thus has little influence in such datasets. However, imputation improves overall workflow robustness and flexibility for different input data types. For example, loss of data from an entire SEC fraction due to failed MS acquisition can robustly be compensated by imputation rather than re-analysis of the fraction or repeat of the entire experiment. Furthermore, missing value imputation should improve the interpretability of datasets affected by more missing values, e.g., when acquired via classical data-dependent MS.

62 Filter peptides by consecutive peptide detection. Peptides that have never been detected in more than N consecutive fractions, here N = 2, are removed from the traces object. This effectively removes false-positive peptide detections from the dataset.

```r
pepTracesConsIds <- filterConsecutiveIdStretches(traces = pepTracesImp,
                                            min_stretch_length = 3,
                                            remove_empty = TRUE)
```

63 Calculate the average SPC for each peptide. For each peptide, the average pairwise correlation with the quantitative traces of its sibling peptides (i.e., peptides derived from the same protein) is calculated (Fig. 3b).

```r
pepTracesSibPepCorr <- calculateSibPepCorr(traces = pepTracesConsIds,
                                          PDF = plotPDF)
```

64 Filter by SPC. Peptides below a minimum average SPC cutoff are removed. The rationale is that outlier peptides and proteins with very heterogeneous quantitative peptide traces are excluded from further analysis. The filtering cutoff can either be automatically determined by a target-decoy–based FDR estimation approach (A, Fig. 3c), or a fixed cutoff can be applied (B).

(A) Target-decoy–based FDR estimation

(i) Make a conservative estimate of the FFT from the unfractionated SEC input sample that was also used to train the PyProphet model for peptide-centric analysis. This is conservative, because we expect to see cumulatively more proteins in the SEC fractions than in the single unfractionated input sample. The estimated pi0 ~ FFT is reported in the protein-level pdf report. For this dataset, the FFT was estimated to be 0.491. Define this in R:

```r
estimatedFFT <- 0.491
```

(ii) Filter by FDR cutoff using the estimated FFT.

```r
pepTraces_filtered_FDR <- filterBySibPepCorr(traces = pepTracesSibPepCorr,
                                             fdr_cutoff = 0.01,
                                             FFT = estimatedFFT,
                                             rm_decoys = TRUE, PDF = plotPDF)
```

! CAUTION This option is valid only if you have continuously kept decoys in your analysis. If you have not done this, then the most conservative strategy is to apply a FFT of 1. However, if you have an FFT estimation available, this will significantly boost your sensitivity and result in a higher number of remaining proteins for the downstream analysis. We have specifically tested this option for datasets processed by an
OpenSWATH-based workflow. If other data processing tools have been used, the decoy-based FDR estimation on SEC level should be treated with caution.

(B) **Absolute SPC cutoff**
(i) Filter by an absolute SPC cutoff, here 0.25.

```
pepTraces_filtered_absoluteCutoff <- filterBySibPepCorr(traces = pepTracesSibPepCorr, 
fdr_cutoff = NULL, 
absolute_spcCutoff = 0.25, 
rm_decoys = TRUE, 
PDF = plotPDF)
```

**Inspect resulting peptide-level traces object**
65 Look at the summary statistics to check the number of detected peptides and proteins.

```
summary(pepTraces_filtered_FDR)
```

66 Plot some example traces to see if the remaining peptides show agreeing intensity profiles along the SEC dimension and to get an impression of the SEC resolution (peak width). Here, we show the exemplary visualization of the proteasome subunit alpha type-1 (UniProt ID = P25786, Fig. 3e).

```
test_protein <- c("P25786")
test_peptide_traces <- subset(traces = pepTraces_filtered_FDR, 
trace_subset_ids = test_protein, 
trace_subset_type = "protein_id")
plot(test_peptide_traces, 
PDF = plotPDF, 
name = paste0("pepTraces_{",test_protein))
```

**Protein quantification**
67 Perform protein quantification by selecting the top N, here N = 2, peptides based on their global intensity across all fractions.

```
protTraces <- proteinQuanti(fication(pepTraces_filtered_FDR, 
topN = 2, keep_less = FALSE)
```

68 Inspect summary statistics of the resulting protein traces.

```
summary(protTraces)
```

Compare the number of remaining proteins to the number of proteins in the peptide-level traces. If the number of proteins is much reduced during the protein quantification step, many proteins might have been detected by a single peptide only.

▲ **CRITICAL STEP** Careful consideration is necessary to decide whether you want to trust such single peptide hits and include them in your downstream analysis by reducing the quantification criteria.

69 Visualize and inspect example protein traces (exemplary visualization of the proteasome subunit alpha type-1 (UniProt ID = P25786)).

```
test_protein_traces <- subset(traces = protTraces, 
trace_subset_ids = test_protein, 
trace_subset_type = "protein_id")
plot(test_protein_traces, colour_by = "Entry_name", 
PDF = plotPDF, 
name = paste0("protTraces_{",test_protein))
```

? **TROUBLESHOOTING**
Overall workflow quality control to evaluate the global proteome assembly state

70 Use the protein-level profiles to estimate the overall complex assembly state observed in the sample as a quality control to ensure the successful extraction and profiling of largely intact complexes. Here, we evaluate the total MS signal in assembled versus monomeric range (Fig. 3d).

```
summarizeMassDistribution(protTraces, PDF = plotPDF)
```

Automatically identify optimal processing parameters based on a protein-level parameter grid search

▲ CRITICAL A grid search can be performed to determine an optimal set of parameters for the protein- and/or complex-centric proteome-profiling workflow. This optimal parameter set depends mostly on the co-fractionation characteristics and MS setup.

71 Randomly select a subset of proteins for the grid search. The selected subset of proteins should be representative of the proteome, thereby providing a trade-off between coverage and computational run-time. From our experience, selecting <100 proteins suffers in regard to robustness, while >500 proteins will require a lot of processing time. We therefore propose a random selection of ~500 proteins.

```
all_proteins <- unique(pepTraces_filtered_absoluteCutoff$trace_annotation$protein_id)
testProtein_idx <- sample(1:length(all_proteins), 500)
testProteins = all_proteins[testProtein_idx]
peptideTracesSubset = subset(traces = pepTraces_filtered_FDR, trace_subset_ids = testProteins, trace_subset_type = "protein_id")
```

72 Perform the parameter grid search. The grid search performs a peptide co-elution peak group finding for a selected combination of parameters with the goal of determining a good parameter set for the following analyses. Please note that the selection of suitable parameters for the grid search is critical.

```
gridFeatures <- performProteinGridSearch(traces = peptideTracesSubset, corrs = c(0.9, 0.95), windows = c(8, 10), smoothing = c(7, 9), rt_heights = c(1, 3), n_cores = 3)
```

▲ CRITICAL The selection of parameters for the grid search is critical. Guidelines for the selection of reasonable parameters are discussed in Box 2.

73 Score protein features across all grid search parameters and select the best parameter set.

```
gridFeatures_scored <- lapply(gridFeatures, calculateCoelutionScore
gridFeatures_qvalues <- lapply(gridFeatures_scored, calculateQvalue, plot = FALSE)
gridFeatures_stats <- qvaluePositivesPlotGrid(
    featuresGrid = gridFeatures_qvalues, colour_parameter = "corr", PDF = plotPDF)
bestParameters <- getBestQvalueParameters(
    stats = gridFeatures_stats, FDR_cutoff = 0.05)
bestParameters
write.table(bestParameters, "bestParameters.tsv", sep = "\t", quote = FALSE, row.names = FALSE)
```

▲ CRITICAL STEP Inspect the pseudo-receiver operating characteristic (ROC) curves generated by the grid search (Fig. 4a). Optimal parameters are at the upper left corner of the
observed distribution. Parameters that are consistently in the upper left corner are especially important.

**Perform protein-centric analysis**

▲ CRITICAL Protein-centric analysis detects peptide co-elution peak groups along the chromatographic dimension. Each detected peak (‘protein feature’) represents the protein in a specific assembly state, i.e., monomeric or bound to different protein complexes.

74 Perform protein feature finding.

```r
proteinFeatures <- findProteinFeatures(traces = pepTraces_filtered_FDR,
corr_cutoff = bestParameters$corr,
window_size = bestParameters$window,
rt_height = bestParameters$rt_height,
smoothing_length = bestParameters$smoothing_length,
collapse_method = "apex_only",
perturb_cutoff = "5%",
parallelized = TRUE,
useRandomDecoyModel = TRUE)
```

75 Score detected protein features and estimate FDR.

```r
proteinFeatures_scored <- scoreFeatures(features = proteinFeatures,
FDR = 0.05,
PDF = plotPDF)
write.table(proteinFeatures_scored,
"proteinFeatures_scored.tsv",
sep = "\t",
quote = FALSE,
row.names = FALSE)
```

▲ CRITICAL STEP Inspect the P value histogram (Fig. 4b,c). There should be a high peak close to zero and a uniform distribution across all other P values.

? TROUBLESHOOTING

76 Inspect summary statistics on resulting protein features. The resulting figures provide information about the number of unique assembly states detected for all the proteins as well as about the number of proteins with at least one assembled protein signal (MW ≥ 2× monomeric MW in SEC) (Fig. 4d).

```r
summarizeFeatures(feature_table = proteinFeatures_scored,
PDF = plotPDF,name = "proteinFeatures_summary")
```

77 Visualize and inspect protein features (Fig. 4c).

```r
plotFeatures(feature_table = proteinFeatures_scored,
traces = pepTraces_filtered_FDR,
calibration = calibration,
feature_id = test_protein,
annotation_label = "Entry_name",
onlyBest = FALSE,
peak_area = TRUE,
monomer_MW = TRUE,
PDF = plotPDF,
name = paste0("protFeatures_",test_protein))
```

78 Plot all detected proteins.

```r
allDetectedProteins <- unique(proteinFeatures_scored$protein_id)
pdf("allDetectedProteins.pdf", height = 6, width = 8)
```
for (protein in allDetectedProteins) {
  plotFeatures(feature_table = proteinFeatures_scored,
               traces = pepTraces_filtered_FDR,
               calibration = calibration,
               feature_id = protein,
               annotation_label = "Entry_name",
               onlyBest = FALSE,
               peak_area = TRUE,
               monomer_MW = TRUE,
               PDF = FALSE) }
  dev.off()

▲ CRITICAL STEP Inspect some detected protein features and evaluate if the detected peak groups correspond to what you would have also selected as peak groups during manual inspection.

? TROUBLESHOOTING

Perform complex-centric analysis

Complex feature-finding represents the central step of complex-centric analysis using CCprofiler. Based on prior protein interaction data and quantitative fractionation profiles, CCprofiler detects groups or subgroups of locally co-eluting proteins, indicating the presence of protein-protein complexes in the biological sample. Target complex queries are supplemented with decoy complex queries to support error control of the reported results. The result is a table summarizing the presence and composition of protein-protein complexes in the biological sample analyzed.

79 Prepare target complex queries. There are two options for protein complex target generation in CCprofiler: use defined protein complex models for direct use as queries (two or more subunits, e.g., from CORUM) (option A) or use a PPI network from which target complex queries can be extracted (option B).

(A) Defined protein complex models

(i) Inspect the coverage of pre-defined protein complex queries from the previously loaded CORUM database (Fig. 5a).

  plotSummarizedMScoverage(hypotheses = corumComplexes,
                           protTraces = protTraces,
                           PDF = plotPDF,
                           name_suffix = "CORUM")

(B) Generate and inspect protein complex queries from binary PPI networks, here based on BioPlex

(i) Calculate pairwise distances between any two proteins in the interaction network.

  pathLengthBioPlexPPIs <- calculatePathlength(BioPlexPPIs)

(ii) Generate protein complex targets by grouping proteins based on a user-defined distance cutoff. Here, we consider only direct neighbors of each protein.

  networkTargetsBioPlexPPIs <- generateComplexTargets(dist_info = pathLengthBioPlexPPIs,
                                                      max_distance = 1,
                                                      redundancy_cutoff = 0)

(iii) Inspect newly generated protein complex queries.

  head(networkTargetsBioPlexPPIs)
  plotSummarizedMScoverage(hypotheses = networkTargetsBioPlexPPIs,
                           protTraces = protTraces,
                           PDF = plotPDF,
                           name_suffix = "BioPlex")
**Critical Step** There are a few points that need to be considered here. First, it is essential that the chosen protein complex queries match the experimental dataset. Therefore, inspect the protein and complex coverage pie charts (Fig. 5a). We recommend that at least half of the proteins and protein complexes represented in the complex query set should be (partially) detected in the experiment. Second, one critical question during complex query generation is how to handle redundancies, i.e., protein complex queries that partially or fully overlap. Due to the complex-centric scoring functions in CCprofiler, we recommend also keeping protein complex subsets in the target queries. Instead of merging/removing overlapping queries at this stage, we recommend collapsing detected complex signals at Step 88. Third, if you are especially interested in some protein complexes that are not present in any available database, you can manually append these complexes to a generated target query list. It is important to keep in mind that the target query list should always contain at least ~1,000 complexes to ensure robust decoy-based FDR estimation and sensitive detection rates. If less complex queries are selected, feature finding can still be performed, but decoy generation and FDR estimation are not applicable.

**Troubleshooting**

80 Prepare decoy complex queries.

```r
correntComplexes <- generateComplexDecoys(target_hypotheses = corumComplexes, 
                                            dist_info = pathLengthCorumComplexes, 
                                            min_distance = 2, 
                                            append = TRUE)
```

**Critical Step** Decoy complex queries are generated based on the target complex query set and its underlying network structure. The minimum distance specifies the minimal number of edges between any two proteins within any generated decoy complex query. It is important that the interaction network based on the targets be large enough to generate a random decoy set that does not overlap with the target complex queries. We recommend complex query sets of ≥1,000 targets for the decoy-based approach.

81 Perform protein complex feature finding. Protein complex features are determined similarly to the protein features described in Step 74. First, CCprofiler applies a sliding window algorithm, where all proteins of a hypothesized protein complex are tested for local profile correlation (i.e., to determine whether all or a subset of these proteins are present in the same SEC range). If a subset of the proteins within a protein complex hypothesis correlate better than the specified cutoff, a protein complex feature is initiated, followed by peak detection within the regions of high correlation.

```r
complexFeatures <- findComplexFeatures(traces = protTraces, 
corem_hypothesis = corumComplexesPlusDecoys, 
corr_cutoff = bestParameters$corr, 
window_size = bestParameters$window, 
rt_height = bestParameters$rt_height, 
smoothing_length = bestParameters$smoothing_length, 
collapse_method = "apex_network", 
perturb_cutoff = "5%", 
parallelized = TRUE, 
n_cores = 3)
```

**Critical Step** If no parameter selection was performed on the protein-centric level, you can also do a complex level grid search.

82 Filter complex features according to their apparent MW, removing protein complex features that elute at an apparent MW lower than any of the monomeric MWs of its subunits.
complexFeaturesFilteredMW <- filterFeatures(
  feature_table = complexFeatures,
  min_monomer_distance_factor = 2)

Select only the best complex feature, i.e., the complex signal with the most subunits and highest correlation. This step is necessary before the statistical scoring, because individual elution peaks are not independent.

complexFeaturesBest <- getBestFeatures(
  feature_table = complexFeaturesFilteredMW)
complexFeaturesBest_scored <- scoreFeatures(
  features = complexFeaturesBest,
  FDR = 0.05,
  PDF = plotPDF,
  name = "complex_qvalueStats")
summarizeFeatures(complexFeaturesBest_scored,
  PDF = plotPDF,
  name = "complexFeaturesBest_feature_summary")

▲ CRITICAL STEP Inspect the P value histogram (Fig. 5b,c). There should be a high peak close to zero and a uniform distribution across all other P values.

? TROUBLESHOOTING

Append secondary features based on a user-defined local subunit correlation cutoff, here 0.5.

complexFeaturesAll <- appendSecondaryComplexFeatures(
  scoredPrimaryFeatures = complexFeaturesBest_scored,
  allFeatures = complexFeaturesFilteredMW, peakCorr_cutoff = 0.5)
write.table(complexFeaturesAll, "complexFeaturesAll.tsv",
  sep = "\t", quote = FALSE, row.names = FALSE)

Inspect summary statistics on resulting protein features (Fig. 5d).

summarizeFeatures(complexFeaturesAll,
  PDF = plotPDF,
  name = "complexFeaturesAll_feature_summary")
plotSummarizedComplexes(
  complexFeatures = complexFeaturesAll,
  hypotheses = corumComplexes,
  protTraces = protTraces,
  PDF = plotPDF)

Visualize and inspect detected complex features (Fig. 5e).

testComplex <- "181"
plotFeatures(feature_table = complexFeaturesAll,
  traces = protTraces,
  calibration = calibration,
  feature_id = testComplex,
  annotation_label = "Entry_name",
  onlyBest = FALSE,
  peak_area = TRUE,
  monomer_MW = TRUE,
  PDF = plotPDF,
  name = paste("complexFeatures_",testComplex))
Plot all detected complexes.

```r
critical step Inspect some detected complex features and evaluate if the detected peak groups correspond to what you would have also selected as peak groups during manual inspection.

**Troubleshooting**

Collapse overlapping and redundant co-elution evidence to delineate complexes and complex families with defined co-elution of subunits in SEC.

```
critical step To retrieve unique, non-redundant protein complex signals, the reported complex signals need to be collapsed based on a strategy that considers (i) subunit composition and (ii) resolution in the chromatographic dimension.

Visualize and inspect all collapsed complex features.

```
Anticipated results

We used the presented protocol to study the proteome assembly state of a population of exponentially growing HEK293 cells\textsuperscript{16}. In this study, we observed 5,124 proteins at 1% protein FDR (Fig. 3c), using the SEC-informed filtering approaches as presented above. We could show that 64% of the proteins are present in at least one assembled state according to the MW distribution along the SEC and that 27% of proteins distribute into multiple distinct assembly states, as evidenced by unique elution peaks (Fig. 4d). Using CORUM as prior information for complex-centric analysis, we could observe evidence for 574 protein complexes at 5% FDR (Fig. 5e), boiling down to 195 unique protein complex signals after feature collapsing. We further demonstrated improved coverage of observable protein complexes by combining prior connectivity information from multiple sources; for example, integrating results based on CORUM, BioPlex and StringDB as prior information detects 462 unique protein complexes\textsuperscript{16}. Furthermore, we demonstrated the sub-complex resolution of protein complex information retrievable by complex-centric proteome profiling. Specifically, the study identified a novel sub-complex of the COP9 signalosome complex (holo-CSN) which, due to the absence of the catalytically active subunit CSN5 and presence of subunits involved in substrate recruitment, may be able to attenuate holo-CSN de-neddylation activities by competitive binding to and sequestration of Cullin-Ring ligase substrate complexes. Second, complex-centric profiling revealed the specific composition and relative abundance of an unexpected, late-stage intermediate of 20S proteasome assembly, the composition of which suggests an alternative sequence of subunit assembly when

Troubleshooting

Troubleshooting advice can be found in Table 5.

| Step | Problem | Possible reason | Solution |
|------|---------|----------------|----------|
| 28   | Missing or poor quality MS data of an isolated chromatographic fraction | Problems in sample workup, NanoLC failure, data corruption | The (re-) analysis of the fraction can be skipped, and intensities can be extrapolated from adjacent fractions that were analyzed successfully. This can be achieved by using the function `imputeMissingValues` of CCpro | |
| 67–69| Error in protein quantification | There are also non-proteotypic peptides in the peptide-level traces | Remove all non-proteotypic peptides (peptides that map to more than one protein) |
| 75   | P value histogram for protein signals does not show a uniform distribution\textsuperscript{53,54} | Too few queries were tested | Check if the number of proteins with \(\geq 2\) peptides is \(\geq 500\) |
| 78 and 87 | Detected protein/protein complex signal apex and boundaries do not look reasonable | Parameters for the feature finding were not selected appropriately | Check the automated parameter grid search again |
| 79 and 80 | Error in target or decoy complex query generation | Selected binary interaction network is too small or not connected | Use matching interaction network for your sample |
| 83   | P value histogram for protein complex signals does not show a uniform distribution\textsuperscript{53,54} | Too few queries were tested | Increase the number of protein complex queries (>1,000 queries) |
|      |         | Too few queries are observable in your dataset | Manually plot the protein profiles of a few standard complexes (e.g., 26S proteasome - CORUM ID 181, CCT complex (chaperonin containing TCP1 complex) - CORUM ID 126) to test whether any good signals can be observed |
|      |         | Error in protein complex decoy generation | Check if the selected parameters are sensible |
|      |         | Check if decoys are present | |

NATURE PROTOCOLS | VOL 15 | AUGUST 2020 | 2341-2386 | www.nature.com/nprot
compared to the canonical model of 20S assembly. Furthermore, the quantitative distribution of proteins across these and other instances of SEC-resolvable protein complexes was assessed.

It can generally be anticipated that complex-centric proteome profiling of human cell lines results in the detection of ≥50,000 proteotypic peptides and 4,000 uniquely detectable proteins (at 1% protein FDR). More than 50% of the protein mass (estimated from total MS signal) is expected to be detectable in the SEC region assigned to the complex-assembled state (appearing with ≥2× the monomeric MW in SEC). Protein-centric detection of co-elution signals from peptide-level chromatograms should yield approximately one to six high-quality protein elution peaks for ≥80% of the detected proteins (q value cutoff = 0.05, equivalent to 5% FDR). In our experience with human cell lines, ~25% of the proteins distribute into multiple distinct assembly states, as evidenced by multiple resolved protein elution peaks along the fractionation dimension. Two-thirds of the proteins are observed at least once eluting in assembled state (peak apex at a MW ≥2× the monomeric MW in SEC). Using the CORUM protein complex database as prior connectivity information for complex-centric analysis in CCprofi ler, human SEC-SWATH-MS data analysis is expected to yield evidence for ~400 protein complexes at 5% FDR, boiling down to evidence for ~200 unique protein complex signals after feature collapsing. Explanations of the result tables exported by CCprofi ler are provided in Table 6 (protein-centric analysis results), Table 7 (complex-centric analysis results) and Table 8 (results of complex feature collapsing).

**Reporting Summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this article.
### Table 7 | Explanation of the reported result table for complex-centric analysis

| Column name                     | Explanation                                                                                                                                 |
|---------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| complex_id                      | Query complex identifier, either from CORUM or assigned based on the seed node used to define the subnetwork query sets (e.g., based on BioPlex or StringDB) |
| complex_name                    | Query complex name, either from CORUM or assigned based on the seed node used to define the subnetwork query sets (e.g., based on BioPlex or StringDB) |
| subunits_annotated              | UniProtKB identifiers of proteins annotated to be in complex based on this query hypothesis                                                                 |
| n_subunits_annotated            | Number of proteins annotated to be in complex based on this query hypothesis complex_id                                                                 |
| subunits_with_signal            | UniProtKB identifiers of proteins that have been detected by MS, subset of the proteins annotated to be in complex based on this query hypothesis complex_id |
| n_subunits_with_signal          | Number of proteins that have been detected by MS, subset of the proteins annotated to be in complex based on this query hypothesis complex_id |
| subunits_detected               | UniProtKB identifiers of proteins that have been detected co-eluting in this complex signal/feature                                                                 |
| n_subunits_detected             | Number of proteins that have been detected co-eluting in this complex signal/feature                                                                 |
| completeness                    | Query hypothesis observation completeness (n_subunits_coeluting/n_subunits_annotated)                                                                 |
| left_sw                         | Preliminary sliding window-based left boundary of the detected co-elution signal/feature                                                                 |
| right_sw                        | Preliminary sliding window-based right boundary of the detected co-elution signal/feature                                                                 |
| sw_score                        | Average Pearson correlation among the protein traces within the preliminary sliding window-based peak boundaries                                      |
| left_pp                         | Left boundary of the detected co-elution signal/feature (after peak picking)                                                                 |
| right_pp                        | Right boundary of the detected co-elution signal/feature (after peak picking)                                                                 |
| apex_mw                         | Estimated MW of the apex of the detected co-elution signal/feature                                                                            |
| area                            | Cumulative subunit MS signal area within the detected co-elution signal/feature (left to right _pp)                                                                 |
| peak_corr                       | Average Pearson correlation among the protein traces within the peak boundaries                                                                  |
| total_intensity                 | Subunit MS signal area within the detected co-elution signal/feature (left to right _pp) for each component subunit detected co-eluting |
| intensity_ratio                 | MS intensity ratios within feature boundaries (based on total_intensity), relative to lowest-intensity component (set to 1)                          |
| stoichiometry_estimated         | Stoichiometry estimate based on rounding the intensity_ratio of each subunit versus the lowest-intensity subunit (of which one copy is assumed to be in the complex). Assumes comparable flyability of different protein’s topN peptides (with associated error). |
| monomer_mw                      | MW of the monomers of the component subunits, information from UniProtKB (mass)                                                                  |
| monomer_sec                     | Predicted elution fraction of monomers, based on the monomer_mw and external size calibration using standard protein’s MW, apex fractions and log-linear regression |
| complex_mw_estimated            | Estimated MW of the detected protein complex signal, given the estimated stoichiometries                                                    |
| complex_sec_estimated           | Estimated SEC fraction number of the detected protein complex signal, given the estimated stoichiometries                                           |
| sec_diff                        | Difference between the estimated SEC fraction number of the detected protein complex signal (complex_sec_estimated) and the actually observed peak apex |
| mw_diff                         | Difference between the estimated MW of the detected protein complex signal (complex_mw_estimated) and the actually observed MW at the peak apex |
| coelution_score                 | Co-elution score, based on peak_corr and corrected for multiple testing within the query group/hypothesis context. NA reported for secondary features since q value estimation and FDR control are performed exclusively on primary/best features (see P value, q value and best_feature). |
| decoy                           | A zero indicates that the complex_id corresponds to a target, ‘NA’ indicates that the reported signal is a secondary feature and a 1 indicates that the reported signal is derived from a decoy protein complex query. |
| P value                         | uncorrected P values along the co-elution score (see BioConductor/qvalue R package). NA reported for secondary features since q value estimation and FDR control are performed exclusively on primary/best features. |
| q value                         | q value as estimated by the Storey-Tibshirani approach (see BioConductor/qvalue R package). Measure of significance similar to the P value, but accounting for the multiple testing problem analogously to the FDR. NA reported for secondary features since q value estimation and FDR control are performed exclusively on primary/best features. |

**Data availability**

The MS data for the HEK293 SEC-SWATH-MS experiment is available at ProteomeXchange Consortium PXD007038 (http://proteomecentral.proteomexchange.org). The R-package CCprofiler is available on GitHub at https://github.com/CCprofiler/CCprofiler/.
Code availability

A detailed protocol on how to perform peptide-centric SEC-SWATH-MS data analysis is available on GitHub at https://github.com/CCprofi/SECSWATH_PeptideCentricAnalysis. A detailed protocol on how to perform complex-centric SEC-SWATH-MS data analysis with the CCprofi package as well as example data of our HEK293 experiment are available on GitHub at https://github.com/CCprofi/SECSWATH_ComplexCentricAnalysis and in the Supplementary CCprofi manual.

Table 8 | Explanation of the reported result table after complex feature collapsing

| Column name             | Explanation                                                                 |
|-------------------------|-----------------------------------------------------------------------------|
| complex_id              | Merged identifier of all complex_ids of the parent queries giving rise to the detection of overlapping, at least partially redundant co-elution signals |
| complex_name            | Merged identifier of all complex_names of the parent hypotheses giving rise to the detection of overlapping, at least partially redundant co-elution signals |
| subunits                | Subunits co-eluting, i.e., components of the integrated signal after collapsing on SEC fraction and composition |
| left_pp                 | Left boundary of the detected co-elution signal/feature; median feature boundary when multiple co-elution signals/features were merged |
| right_pp                | Right boundary of the detected co-elution signal/feature; median feature boundary when multiple co-elution signals/features were merged |
| apex                    | Apex of the detected co-elution signal/feature; median of apexes when multiple co-elution signals/features were merged |
| unique_feature_identifier | Unique identifier of the given co-elution signal/feature                      |

References

1. Bludau, I. & Aebersold, R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. Nat. Rev. Mol. Biol. 21, 327–340 (2020).
2. Huttlin, E. L. et al. The BioPlex network: a systematic exploration of the human interactome. Cell 162, 425–440 (2015).
3. Huttlin, E. L. et al. Architecture of the human interactome defines protein communities and disease networks. Nature 545, 505 (2017).
4. Hein, M. Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163, 712–723 (2015).
5. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. J. Cell Biol. 196, 801–810 (2012).
6. Liu, X., Yang, W., Gao, Q. & Regnier, F. Toward chromatographic analysis of interacting protein networks. J. Chromatogr. A 1178, 24–32 (2008).
7. Dong, M. et al. A “tagless” strategy for identification of stable protein complexes genome-wide by multidimensional orthogonal chromatographic separation and iTRAQ reagent tracking. J. Proteome Res. 7, 1836–1849 (2008).
8. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. Nat. Methods 9, 907 (2012).
9. Kristensen, A. R. & Foster, L. J. Protein correlation profiling-SILAC to study protein-protein interactions. in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). Methods in Molecular Biology (Methods and Protocols) Vol. 1188 (ed. Warscheid, B.) 263–270 (Humana Press, 2014).
10. Havigimana, P. C. et al. A census of human soluble protein complexes. Cell 150, 1068–1081 (2012).
11. Panorama of ancient metazoan macromolecular complexes. Nature 525, 339–344 (2015).
12. Kirkwood, K. J., Ahmad, Y., Larance, M. & Lamond, A. I. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. Mol. Cell. Proteomics 12, 3851–3873 (2013).
13. Larance, M. et al. Global membrane protein interactome analysis using in vivo crosslinking and mass spectrometry-based protein correlation profiling. Mol. Cell. Proteomics 15, 2476–2490 (2016).
14. Scott, N. E. et al. Interactome disassembly during apoptosis occurs independent of caspase cleavage. Mol. Syst. Biol. 13, 906 (2017).
15. Stacey, R. G., Skinnider, M. A., Scott, N. E. & Foster, L. J. A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). BMC Bioinformatics 18, 457 (2017).
16. Heusel, M. et al. Complex-centric proteome profiling by SEC-SWATH-MS. Mol. Syst. Biol. 15, e8438 (2019).
17. Scott, N. E., Brown, L. M., Kristensen, A. R. & Foster, L. J. Development of a computational framework for the analysis of protein correlation profiling and spatial proteomics experiments. *J. Proteomics* **118**, 112–129 (2015).
18. Heusel, M. et al. A global screen for assembly state changes of the mitotic proteome by SEC-SWATH-MS. *Cell Syst* **10**, 133–155.e6 (2020).
19. Pauling, L., Itano, H. A., Singer, S. J. & Wells, I. C. Sickle cell anemia, a molecular disease. *Science* **110**, 543–548 (1949).
20. Bache, N. et al. A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
21. Wessels, H. J. C. T. et al. LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. *Proteomics* **9**, 4221–4228 (2009).
22. Ong, S. E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
23. Hu, L. Z. et al. EPIC: software toolkit for elution profile-based inference of protein complexes. *Nat. Methods* **16**, 737–742 (2019).
24. Glatter, T., Wepf, A., Aebersold, R. & Gstaiger, M. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.* **5**, 237 (2009).
25. Roncagalli, R. et al. Quantitative proteomics analysis of signaling dynamics in primary T cells identifies the surface receptor CD6 as a Lat adaptor–independent TCR signaling hub. *Nat. Immunol.* **15**, 384–392 (2014).
26. Collins, B. C. et al. Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 13-33 system. *Nat. Methods* **10**, 1246–1253 (2013).
27. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
28. Picotti, P. & Aebersold, R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).
29. Schubert, O. T. et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).
30. Collins, B. C. et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).
31. Bruderer, R. et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics* **16**, 2296–2309 (2017).
32. Kelstrup, C. D. et al. Performance evaluation of the Q exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
33. Meier, F. et al. Parallel accumulation—serial fragmentation combined with data-independent acquisition (diaPASEF): bottom-up proteomics with near optimal ion usage. Preprint at https://www.biorxiv.org/content/10.1101/656207v2 (2019).
34. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031 (2014).
35. Picotti, P. et al. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494**, 266–270 (2013).
36. Blattmann, P. et al. Generation of a zebrafish SWATH-MS spectral library to quantify 10,000 proteins. *Sci. Data* **6**, 190011 (2019).
37. Heusel, M. Complex-Centric Proteome Profiling by SEC-SWATH Mass Spectrometry. Dissertation, ETH Zurich (2017). https://www.research-collection.ethz.ch/handle/20.500.11850/220300
38. 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).
39. Röst, H. L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32**, 219 (2014).
40. Reiter, L. et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **8**, 430–435 (2011).
41. Teleman, J. et al. DIANA—algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* **31**, 555–562 (2015).
42. Rosenberger, G. et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods* **14**, 921–927 (2017).
43. Röst, H. L. et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods* **13**, 777 (2016).
44. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2009).
45. Franceschini, A. et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2012).
46. Sklarczyk, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
47. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
48. Choi, S. G., Richardson, A., Lambourne, L., Hill, D. E. & Vidal, M. Protein interactomics by two-hybrid methods. *Methods Mol. Biol.* **1794**, 1–14 (2018).
49. Gavin, A.-C. et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
50. Krogan, N. J. et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
51. Guruharsha, K. G. et al. A protein complex network of *Drosophila melanogaster*. *Cell* **147**, 690–703 (2011).
52. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
53. Burger, T. Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *J. Proteome Res.* **17**, 12–22 (2018).
54. Breheny, P., Stromberg, A. & Lambert, J. p-value histograms: inference and diagnostics. *High Throughput* **7**, E23 (2018).
55. Adusumilli, R. & Mallick, P. Data conversion with ProteoWizard msConvert. in *Proteomics. Methods in Molecular Biology* Vol. 1550 (eds. Comai, L., Katz, J. & Mallick, P.) 339–368 (Humana Press, 2017).
56. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
57. Hirano, Y. et al. A heterodimeric complex that promotes the assembly of mammalian 20S proteasomes. *Nature* **437**, 1381–1385 (2005).
58. Hirano, Y. et al. Dissecting β-ring assembly pathway of the mammalian 20S proteasome. *EMBO J.* **27**, 2204–2213 (2008).

**Acknowledgements**

The project was supported by the SystemsX.ch projects PhosphoNetX PPM and project TbX (to R.A.) and the European Research Council (ERC-20140AdG 670821 to R.A.). M.H. was supported by a grant from Institut Mérieux. I.B. was supported by the Swiss National Science Foundation (grant no. 31003A_166435). B.C.C. was supported by a Swiss National Science Foundation Ambizione grant (PZ00P3_161435). A.B.-E. was supported by the National Institutes of Health project Omics4TB Disease Progression (U19 AI106761).

**Author contributions**

I.B., M.H. and R.A. wrote the manuscript with input from all authors. I.B. and M.H. developed the presented workflow, implemented the analysis scripts and performed all analyses. M.H. developed and optimized the experimental protocol for SEC-SWATH-MS. G.R. and M.H. optimized the peptide-centric analysis for SEC-SWATH-MS applications. I.B., M.H., M.F., G.R., R.H. and A.B.-E. developed the CCprofi ler software. A.V.D. performed validation experiments. R.A., M.G., B.C.C. and M.H. conceptualized the primary study. B.C.C., M.G. and R.A. supervised the study.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Extended data is available for this paper at https://doi.org/10.1038/s41596-020-0332-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-020-0332-6.

Correspondence and requests for materials should be addressed to M.G. or R.A.

Peer review information *Nature Protocols* thanks Bruno Manadas, Leonard Foster and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 6 October 2019; Accepted: 17 April 2020; Published online: 20 July 2020
Extended Data Fig. 1 | Overlap of proteins and protein complexes across the SEC fractionation dimension. a, Heat-map representation of the percentage of detected proteins that are shared between each pair of SEC fractions (top). The percentage overlap is calculated as the number of shared proteins relative to the total set of proteins detected in any pair of SEC fractions, as a percentage. The bottom panel illustrates the average percentage of overlapping proteins at different distance thresholds between SEC fractions. b, Heat-map representation of the percentage of detected protein complexes that are shared between each pair of SEC fractions (top). The percentage overlap is calculated as the number of shared protein complexes relative to the total set of protein complexes detected in any pair of SEC fractions, as a percentage. The bottom panel illustrates the average percentage of overlapping protein complexes at different distance thresholds between SEC fractions.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
ProteoWizard (http://proteowizard.sourceforge.net/)
Docker (e.g. https://docs.docker.com/docker-for-windows/)
OpenSwath docker pipeline (See http://openswath.org/en/latest/docs/docker.html)
R (v3.6.0, https://cran.r-project.org/bin/windows/base/) with packages devtools, data.table, ggplot2 and CCprofiler (See setup)
R-package CCprofiler available on GitHub at https://github.com/CCprofiler/CCprofiler/

Data analysis
A detailed protocol on how to perform peptide-centric SEC-SWATH-MS data analysis is available on GitHub at https://github.com/CCprofiler/CCprofiler/SEC_SWATH_PeptideCentricAnalysis
A detailed protocol on how to perform complex-centric SEC-SWATH-MS data analysis with the CCprofiler package as well as example data of our HEK293 experiment are available on GitHub at https://github.com/CCprofiler/SEC_SWATH_ComplexCentricAnalysis and in the supplementary CCprofiler manual.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The mass spectrometry data for the HEK293 SEC-SWATH-MS experiment is available at ProteomeXchange Consortium PXD007038 (http://proteomecentral.proteomexchange.org)
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | 1 replicate |
|-------------|-------------|
| Data exclusions | no data was excluded from the analysis |
| Replication | Reproducibility was examined and confirmed in the primary research publication by Heusel and Bludau et al. (2019) |
| Randomization | not applicable in a single replicate |
| Blinding | not applicable |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- n/a
- □ Antibodies
- □ Eukaryotic cell lines
- □ Palaeontology
- □ Animals and other organisms
- □ Human research participants
- □ Clinical data

### Methods

- n/a
- □ Involved in the study
- □ ChIP-seq
- □ Flow cytometry
- □ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)  

| Cell line source(s) | HEK293 cell line (American Type Culture Collection Cat. No. CRL-1573). |
|---------------------|---------------------------------------------------------------------|
| Authentication      | Hek293 cells were authenticated visually under the microscope, no molecular authentication was performed. |
| Mycoplasma contamination | All cell lines tested negative for Mycoplasma contamination |

Commonly misidentified lines

(See [ICLAC register](#))

| Commonly misidentified lines | NA |
|------------------------------|----|