Adventures in Molecular Wonderland: Exploring the Complexity of Protein Complexes Through Complexome Profiling - And the Role of Size-Exclusion Chromatography

---

Writing Assignment
Molecular and Cellular Life Sciences

Author
Marèl F.M. Spoelstra, BSc
(6101909)

Examiners
Dr. Joost Snijder
Department of Chemistry – Biomolecular Mass Spectrometry & Proteomics

Dr. Kelly E. Stecker
Department of Pharmaceutical Sciences – Biomolecular Mass Spectrometry & Proteomics

---

Educational Institution
Utrecht University

Date
06-03-2024

Layman's Summary

To be able to understand how a cell functions, one has to look at the main effectors, which are the proteins. However, understanding how they work, could not only be solved by deciphering the set of functions for a single protein, as almost all proteins rely for their function on the interaction with others. Several proteins noncovalently bind to each other, forming a protein complex, whereby a specific protein can be part of several different complexes, depending on the requirement of the cell at a specific moment. The formation and disassembly of these complexes are regulated by important cell signaling pathways, for instance by phosphorylation.

When we want to study these protein complexes, we have to purify them from cells and try to keep them intact. Therefore, purification has to be compatible with buffers that don't disrupt the complexes. One of these methods is Size-Exclusion Chromatography (SEC). This technique separates all the different complexes by size into separate fractions, making it now possible to study which proteins are part of which specific complex. SEC is already a widely used method to measure protein complex formation, and by coupling it to mass spectrometry (MS), it allows us to monitor protein complex formation on a systems-wide scale.

To tell which proteins are present, the different fractions are analyzed by MS. The different proteins are enzymatically digested into smaller pieces, the peptides, of which the specific mass could be determined. By fragmenting the peptides into their building blocks, the amino acids, sequences could be retrieved. This information can be seen as a fingerprint of the protein and allows us to trace back which proteins were present in each fraction. Together with the SEC-data, this elucidates which protein complexes were present.

The strengths of this technique, but also the remaining experimental and technical caveats that have to be addressed, are discussed in this review. For example, data analysis becomes more difficult with this amount of data produced, whereby the method should also be able to tell which proteins are truly interacting with each other, or which coincidentally have approximately the same mass and are therefore separated into the same fraction by SEC. Differences in the amount of a complex could be due to changes in the level of expression of a protein, but also due to a change of distribution of this protein over several subunits, which data-analysis tools should also be able to address. Also, the phosphate groups that are important regulators, are hard to find, because they are less abundant and might be missed, for which more specialized approaches are developed.

Tools like CCprofiler allow for error control to identify true complexes, but also several machine-learning-based tools are developed to predict novel protein complexes. Furthermore, several tools and methods are developed to integrate information about the cell signaling pathways, like protein phosphorylation states. With COPF it is now possible to find specific phosphorylation patterns, belonging to either a specific assembly state of the complex or to a specific cellular condition in which this occurs. With the ongoing process in machine-learning approaches, further optimizations in the sensitivity of SEC and MS, and the increasing amount of information accessible in databases, but also the integration of additional purification steps, it will become less challenging to study these complexes.

(544 words)

Adventures in Molecular Wonderland: Exploring the Complexity of Protein Complexes Through Complexome Profiling - And the Role of Size-Exclusion Chromatography

March 6th 2024

Marèl F.M. Spoelstra*

Bijvoet Centre for Biomolecular Research –Biomolecular Mass Spectrometry and Proteomics, Utrecht University, Utrecht, The Netherlands

Keywords: Complexome Profiling, Correlation Profiling, SEC, SEC-SWATH-MS, CCprofiler, ComplexFinder, COPF

indicate the number of words: 9672

ABSTRACT Proteins do not act on their own but form different complexes in the cell, to exert their biological function. This dynamic process of protein complex assembly is fine-tuned, in part by post-translational modifications (PTMs), like phosphorylation. Elucidating the different complexes of a cell, the complexome, and studying these upon perturbations, allows us to understand cellular functions and how they are interrupted in certain conditions or diseases, in a systems-wide view. Complexome profiling poses several analytical challenges, for which sophisticated methods have been developed to purify and analyze these protein complex mixtures. These methods lean heavily on Mass Spectrometry (MS), as it can handle complex samples, with relatively high sensitivity and in a high-throughput manner. The protein complexes are obtained by targeted approaches, like IP-MS, AP-MS, and Proximity Labelling, or by untargeted approaches, based on the biochemical separation of the complexes, like Size-Exclusion Chromatography (SEC). SEC purifies complexes from native conditions, in a high-throughput manner, with a smaller risk of interfering with protein function compared to targeted approaches. As complexome profiling produces a lot of data, several tools exist to analyze this systematically, additionally providing validation tools to minimize the risk of false positives. This review presents some recent literature on SEC-based complexome profiling, which data-analysis toolkits have been developed, how assembly- or condition-specific PTMs could be studied, which caveats these approaches still possess, and which further improvements are being made or what should be an area of interest for follow-up studies.

INTRODUCTION

Understanding how cells work, and most importantly in the context of certain conditions like disease, can be achieved at different levels. We can look at the DNA-level, but also at the level of the transcriptome to decipher which genes are expressed at a certain moment. However, this does not tell us the complete story of the cell, as the main effectors are the proteins that are expressed. This field is tackled by proteomics, whereby the proteins, and the different proteoforms, present in the cells are identified and quantified, which is used to study protein function and regulation in different conditions or even to study changes over time (Altelaar et al., 2013; Ludvigsen & Honoré, 2018; Mann et al., 2013). Although this proteomics approach gives meaningful information about cell function in certain specified conditions, the reality is even more complex. Most proteins do not act on their own but are part of a larger protein complex. Protein function is related to the proteins or other molecules they interact with or the complexes they are part of. This is dynamic, and proteins can interact with different proteins and complexes, depending on the demands and environment of the cell at a specific moment, which is also referred to as 'modular biology'. According to this description, several functional 'modules', the protein complexes, make up the cellular organization, and each of these has a discrete function, due to either spatial separation, or a chemical specificity not shared by the other modules. For some cellular processes, these modules must be insulated from others, to prevent interfering interactions, while for other processes they must be connected to allow for the integration of information inside the cell (Aebersold & Mann, 2016; Altelaar et al., 2013; Budayeva & Cristea, 2014; Hartwell, 1999). It is estimated that most of the proteins are present in an assembled state inside the cell (Bludau et al., 2023; Heusel et al., 2019). These complexes are fine-tuned to perform specific tasks inside the cell, and the formation and disassembly of the different subunits must be tightly controlled. Some proteins might only be able to properly function while they are assembled in a specific complex, while other proteins can be part of several complexes, whereby their function changes when they partake in a different complex (Budayeva & Cristea, 2014).

One example of this is RIPK1, which can be found in different types of complexes, which can either be pro-cell-survival or pro-cell-death complexes. Cell signaling can determine which complex is formed, and the involved proteins determine if either the kinase domain of RIPK1 is active or not, deciding between either cell survival or cell death by apoptosis, or necroptosis (Clucas & Meier, 2023). This example nicely illustrates that studying cell function on only a proteome level is not enough to explain how certain tasks inside the cell are performed or why. Studying the complexome, also known as the interactome, i.e. all of the protein complexes of the cell or organism, is

important to gain more insight into cell function. This has already been studied quite extensively for mitochondrial proteins, and the applied research methodology might be adapted to complexome profiling in other biological systems.

Essential for complexome profiling is the field of mass spectrometry (MS). This technique is excellent for obtaining high-throughput data, with relatively high sensitivity, to identify and quantify proteins in complex biological matrices. Furthermore, relatively low sample input is needed, and even complex samples can be studied by mass spectrometry. In addition, multiplexing methods exist, allowing for parallel quantitative measurement of several experiments or conditions. Specialized techniques are invented with which additional structural information from the MS data can be obtained, like cross-linking MS (XL-MS).

Proteins can be measured at the protein level, referred to as native MS, or at the peptide level, which is known as bottom-up proteomics. For bottom-up approaches, the proteins are digested by proteases, most commonly Trypsin, to obtain peptides, which are separated by reversed-phase chromatography, most often using a C18 column, after which they are ionized and injected into the mass spectrometer. The intact peptides are measured (MS1), after which they are fragmented to obtain information about the amino acid sequence (MS2). With this information, the proteins could be identified and quantified. Different measurement approaches exist, which can be summarized into either Data-Dependent Acquisition (DDA) or Data-Independent Acquisition (DIA). In DDA, a select number of precursor peptides are selected for fragmentation and measurement in MS2, based on for example the top highest-intensity peptides, whereas in DIA all precursors are selected for fragmentation (Figure 1). Therefore, DIA is overall more sensitive as also lesser abundant proteins are more intensively measured. On the other hand, the spectra from DIA measurements are much more complicated to interpret and require specialized analysis tools and a predefined spectral library containing both chromatographic and mass spectrometric information about the peptides, which is needed to deconvolute the signals (Krasny & Huang, 2021).

Before the samples are measured on LC-MS/MS, the protein complexes must be purified and separated. Nowadays, several techniques are used to achieve this, whereby some techniques use targeted approaches, where prior information about the proteins of interest is necessary to design the experiments, while other techniques are untargeted. Targeted approaches allow for the enrichment of a specific complex and its protein constituents, using experimental strategies to pull out possible interaction partners of the target protein by adding a 'handle' on the protein of interest, with which it can be fished out. These are limited by the available antibodies, used for Immuno-Precipitation MS (IP-MS) or the possibilities to use a tag or enzyme on the protein of interest, which should be ectopically expressed for Affinity-Purification MS (AP-Ms), or Proximity Assays (PA), respectively. In untargeted methods, protein complexes are fractionated first, which can be performed by a set of different techniques, based on different biochemical properties. These techniques are formerly referred to as protein correlation profiling. Separation can be achieved by Blue-Native (BN) PAGE gel filtration, Density Gradient Centrifugation, or Chromatography separation techniques, most often Size Exclusion Chromatography (SEC) (Cabrera-Orefice et al., 2022; Low et al., 2021; Smits & Vermeulen, 2016).

With this literature review, we want to elucidate how complexes can be studied, and which information can be obtained by different types of experiments, like differences in assembly states upon perturbations, possible integration of spatial information, and proteoform-specific complex states. The focus will mainly be on SEC-based complexome profiling, as this technique is suitable for the study of a large set of complexes. It does not require ectopic expression of protein constructs, using either a tag or enzyme, making it also suitable for studying patient-derived input materials. Moreover, it is already being commonly applied in an MS pipeline, for which several high-throughput methods and systematic data-analysis tools have been developed. The main advantages of SEC-based correlation profiling, the common limitations, recent advances, and possible improvements for future research will be discussed. Some highly promising data-analysis tools that enable handling the amount of data common for these types of experiments, in a more systematic manner, as well as providing validation or error-controls for these experiments, are reviewed here.

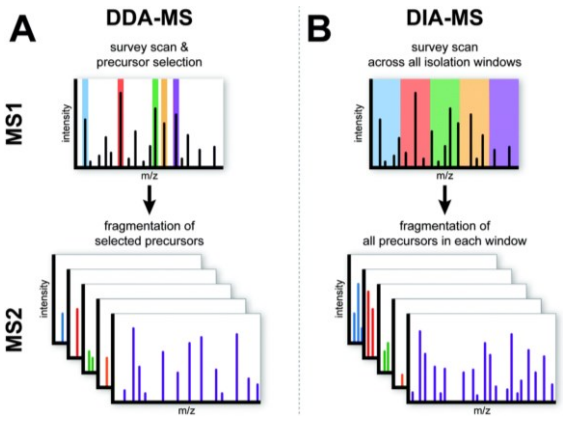

Figure 1: Figure from Krasny & Huang, 2021. In DDA-MS, precursor-ion selection for fragmentation and MS2 analysis is based on for example the top abundant peptides of the MS1 scan. In DIA-MS, a selection window with a predefined size selects all precursor-ions within, resulting in the fragmentation and analysis of all precursor-ions, including lower abundant peptides.

COMPLEXOME PROFILING – *Targeted Strategies*

Targeted approaches make use of co-purification of proteins that are part of the same complex.
In IP-MS methods, an antibody specific for a protein of interest is conjugated to beads, with which this protein and its interactors are concurrently purified. This technique is limited by the available antibodies. As a benefit, native physiological conditions can be used to study the protein-protein interactions, as there is no need for the expression of specifically designed protein constructs. However, possible interaction partners could be missed by this technique when the binding epitope coincides with an important binding domain, but also unspecific binding proteins might be co-purified (Figure 2) (Gnanasekaran, 2023; Low et al., 2021; Smits & Vermeulen, 2016).

AP-MS relies on the same principle, by tagging a target protein and simultaneously pulling out the interaction partners. A specific tagged protein construct is designed and expressed, using for example a Strep-tag, a GFP-tag, or a FLAG-tag, which has affinity for specific beads and is then used to purify the complexes. This technique circumvents possible limits such as low endogenous expression of the bait-protein in IP-MS, due to higher ectopic expression of the protein. On the other hand, the higher expression might not resemble real native physiological conditions, and tags might interfere with protein function, by for example changing binding affinities or subcellular localization (Figure 2) (Low et al., 2021; Morris et al., 2014).

Instead of tagging a bait protein with a purification tag, an enzyme could be fused to it, which is used in proximity-based approaches, like for example BioID. The enzyme that is linked to the protein will catalyze the addition of specific modifications on the proteins that are in close vicinity of the bait protein. For example in BioID where a biotin ligase is added, decorating the target proteins with biotin moieties, which can afterwards be purified by affinity purification using streptavidin beads. Due to the specific distance restraints of the enzyme, these types of techniques allow for high spatial resolution. Furthermore, some enzymes might react relatively fast, in minutes, while others take several hours, thereby allowing for fine-tuning on which time-scale interactions are studied, but also allowing for high temporal resolution (Kong et al., 2022; Low et al., 2021).

All of these targeted approaches have a relatively high risk of the co-purification of unspecific proteins, and several control samples or control steps during data analysis ought to be added, to determine which proteins are true and false positives. Furthermore, it can be difficult to discriminate between direct binders, indirect binders, and other proximal proteins (Figure 2) (Kong et al., 2022; Low et al., 2021; Smits & Vermeulen, 2016).

COMPLEXOME PROFILING – *Correlation Profiling*

Untargeted strategies rely on co-fractionation or co-elution techniques, also known as correlation profiling, whereby the protein complexes are fractionated under native conditions, based on their biophysical properties, assuming that proteins of one complex have similar migration or elution profiles (Figure 2). A great benefit of these approaches is that the whole complexome is studied simultaneously, whereas targeted approaches only look at specific complexes one at a time. Furthermore, compared to affinity purification MS (AP-MS), where an antibody is used to select a protein and its direct interaction partners, these methods increase the chance that also some indirect interactions can be found, as these will still co-elute, but might not have strong enough binding to be elucidated with AP-MS.

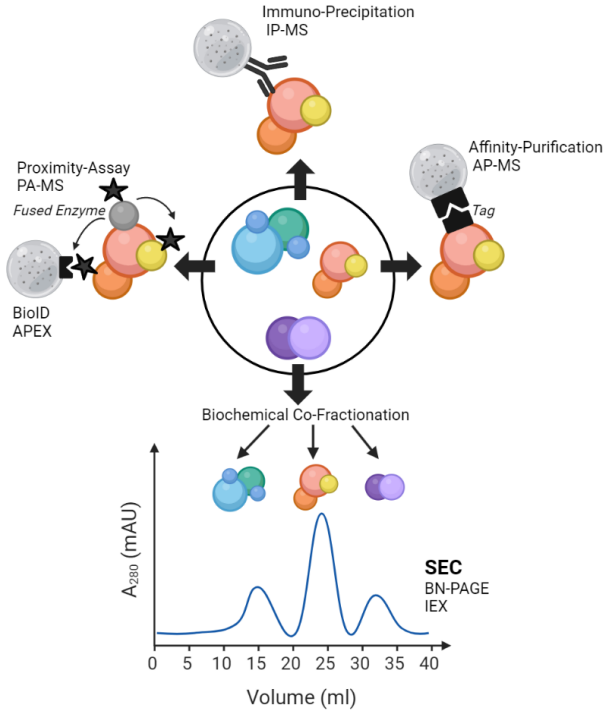

Figure 2: Adapted from Smits & Vermeulen, 2016. Schematic drawing of the different techniques that are commonly used to study the complexome. In Immuno-Precipitation, an antibody fused to a bead targets the bait protein and is used to purify this protein and its interactors. In AP-MS, instead of an antibody, the bait protein is ectopically expressed with a tag, that has affinity for a ligand that is immobilized on stationary material. For Proximity-Assays, an enzyme is fused to the bait protein, which enzymatically decorates the neighboring and interacting proteins, like for example biotin for BioID, which can be used for purification, like for example streptavidin-beads in the case of BioID. Biochemical Fractionation makes use of the principle that proteins of the same complex will separate together, showing similar elution or migration profiles. Created with BioRender.com.

Proximity-based approaches also allow for studying both direct and indirect interactions.

One of the oldest methods is using a density gradient whereby protein complexes are separated based on their sedimentation rates, like in a sucrose gradient. These gradients suffer from poor resolution, due to a relatively large spread of the proteins over the gradient. Also, for this technique, it can be more difficult to determine whether the co-migration is due to the protein being within the same complex, or having similar properties leading to similar sedimentation rates (Cabrera-Orefice et al., 2022; Salas et al., 2020).

Another well-known method relies on separation by size and electrophoretic mobility using gel electrophoresis, called Blue Native-polyacrylamide gel electrophoresis (BN-PAGE), as native buffers are used, as well as using blue staining of the proteins. This method allows for higher resolution, uses relatively low amounts of input material (on the microgram scale), and can also be used to study membrane proteins (Cabrera-Orefice et al., 2022; Salas et al., 2020; Wittig et al., 2006). Furthermore, this technique is typically suitable for separating proteins from 0.02MDa to 10MDa, although specialized gels with large pores are used to separate proteins up to 45MDa. These gels are made by increasing the concentrations of the crosslinker bis-acrylamide (Strecker et al., 2010). Gel electrophoresis can separate the complexes with high resolving power, but to be able to retrieve this resolution in the following steps, a sufficient amount of slices should be picked from the gel. In general, up to 70 slices are made per gel-lane.

In order to increase the resolution even further, to be able to also find less abundant protein complexes, super-complexes, and the subunit composition, a novel method was developed whereby the gels were frozen, and sliced by cryo-microtome slicing. For their method, they were able to slice the gel into 230 slices, with a slice size of 0.3mm, compared to 1mm slices when done manually. The slices were digested and analyzed with high-performance LC-MS/MS using Label-Free Quantification (LFQ). The peptide intensities are then integrated over time to provide Peak Volumes, and a mass-calibration step and elution time shift correction were applied, which is important as 230 datasets have to be combined. The peptide PV profiles are normalized across the datasets, and the relative protein abundance profile could be calculated from the average of at least two protein-specific profiles and normalized for their abundance. Based on the correlation of peptide and protein profiles, the complexes were identified. To determine the apparent molecular mass for each protein complex, reference protein complex peaks with known masses were also measured and a linear regression of their slice numbers was used and applied to the experimental slice number of the protein complex. The Full-Width at Half Maximum (FWHM) indicates the effective resolution that was achieved, which was on average 1mm of the gel, corresponding to 3 slices, for the best-resolved peaks (Muller et al., 2016; Müller et al., 2019).
BN-PAGE methods can also be used with multiplexing approaches, to simultaneously measure different conditions in one MS run (Guerrero-Castillo et al., 2021).

SEC-BASED COMPLEXOME PROFILING

Other co-fractionation methods rely on chromatography-based separation techniques. For this review, the focus is more in-depth on Size-Exclusion Chromatography (SEC) approaches as a prefractionation method, which is already commonly applied in combination with LC-MS/MS. The separation of molecules in SEC is based on the hydrodynamic radius of the protein or protein complex. The stationary phase consists of porous beads, whereby smaller particles can access the pore of the beads and see the whole column volume, whereas the larger particles are less likely to enter all pores and spend less time in the pore volume. Therefore, the larger molecules can migrate faster through the column and will elute first, and the smallest molecules elute last (Burgess, 2018).

SEC is suitable for the purification of a large range of complex sizes due to the choice of several types of resin with ranging bead and pore sizes that are available, e.g. up to sizes of 40MDa for Sepharose resins. In addition, it can be automated for high-throughput approaches and can provide some additional information about the complexome using specialized software that was developed recently.

In addition, SEC is highly suitable, as native buffers are compatible with this separation approach, in contrast to for example reversed-phase, which requires organic buffers that could denature proteins and disassemble the protein complexes. On the other hand, the resolution is a bit lower for SEC, compared to the other chromatography methods. The resolution in chromatography describes the ability to separate two peaks based on the difference in retention times and the corresponding peak widths. The resolution depends therefore on the selectivity of the resin, i.e. the differences between retention times, and the efficiency with which it can produce narrow peaks. The efficiency becomes higher when smaller beads are used. The peak capacity describes the maximum theoretical number of peaks that can be separated on the column, by dividing the gradient time by the average peak width. For example, ion exchange chromatography (IEC) methods have stronger interactions with the column, enhancing the retention and thus showing smaller profile widths, thereby having a higher resolution compared to SEC. When the same gradient time is used, IEC would also have a higher peak capacity than SEC.

To keep the complexes stable, the columns are run at lower temperatures, which increases separation time, compared to separations at higher temperatures. Important for the separation resolution is the column dimension that is used, e.g. column diameter and lengths, as well as the beads that are chosen. This also

determines the minimal number of fractions that should be collected, to retrieve this resolution, like the number of slices that should be chosen for BN-PAGE methods. A trade-off has to be made between the fractionation range, the range of molecular weights that have access to the pore, and the resolution, as larger beads allow for sufficient separation of larger protein complexes, but also lower the resolution that can be achieved. For a high-throughput and high-resolution SEC-SWATH-MS method, the optimal bead size was considered to be 3µm with 500Å pores, using a 300x7.8 mm column (Bludau et al., 2020; Cabrera-Orefice et al., 2022; Heusel et al., 2019; Salas et al., 2020).
Protein elution profiles can correlate due to the proteins being within the same complexes, however, these results should also be treated with caution because it is still likely that co-elution occurs due to proteins having similar biophysical properties on which the separation was based. The amount of false positives could be limited by using high-resolution separation, as well as using an orthogonal separation approach.

Additionally, complexome studies produce a lot of data, containing the SEC protein elution profiles, and MS/MS information obtained at the peptide level, like quantitative values, peptide sequences, and the corresponding protein groups. Many proteins are identified in one MS run, which should be correlated and compared to the migration or elution profiles of the SEC fractions, to identify the complexes and their protein subunits. Several methods have been developed to be able to do this in a high-throughput and automated manner, like COPAL, ComPrAn, ComplexFinder, and CCprofiler.
COPAL is developed to compare different BN-PAGE gels by using a 'multiple gel alignment' with which experimental replicates could be compared, and this takes biological and technical variations into account, like differences in overall migration pattern, enabling identification of true variations caused by mutations, or other perturbations. A gene set enrichment is used to select protein complexes mostly affected by mutations or other changes in condition (Van Strien et al., 2019).
ComPrAn is developed for the analysis of complexome data from density gradient studies, but can also be applied to other chromatography methods, which use heavy/light labeling for two conditions, like SILAC (Páleníková et al., 2021).
We will mainly focus on CCprofiler, as this one is especially suitable for SEC data, although it could be used for other fractionation methods, and provides an error control to limit the number of false-positives. CCprofiler is a complex-centric method, developed for SEC-SWATH-MS, which includes prior complexome information, and allows for higher sensitivity measurements, and better signal-to-noise, due to filtering possibilities, as well as validation for the complexes found (Bludau et al., 2020; Cabrera-Orefice et al., 2022; Heusel et al., 2019; Salas et al., 2020). The higher sensitivity of this method is mostly due to the use of Data Independent Acquirement (DIA), i.e. Sequential Window Acquisition of All Theoretical Proteins (SWATH), as this also measures less abundant peptides, compared to DDA, which focuses mainly on higher abundant peptides (Krasny & Huang, 2021).

This method was applied to a HEK293 cell lysate, but optimizations could be made for other input material, or complexes of interest, as this resin, a Yarra 3-µm SEC-4000 column, has a fractionation range up to 1.5MDa, and might not capture larger complexes (Bludau et al., 2020, 2023; Heusel et al., 2019).

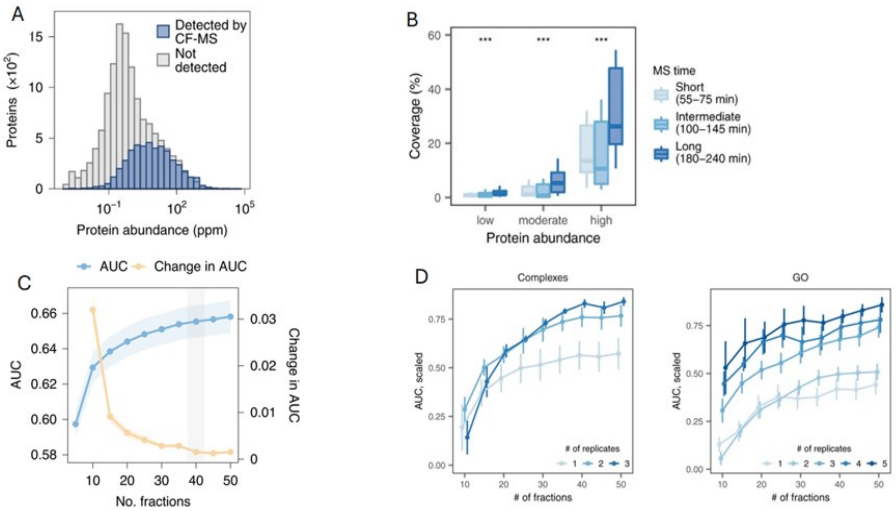

Figure 3: Figure from Skinnider & Foster, 2021. In A) is shown that proteins with higher abundance are better covered by most Co-Fractionation-MS (CF-MS) methods, shown for human proteins. B) Using longer gradient times shows higher coverage of lower abundant proteins C) The AUC indicates the precision with which known complexes are identified from the CF-MS data, with 0.5 being random performance. Collecting more than 40 fractions per replicate did not result in large improvements in the performance. D) Using more replicates per sample has a large impact on the AUC, shown for both the AUC based on the recovery of known complexes, and proteins with the same GO-term.

A recent meta-analysis study, which considered 200 co-fractionation MS studies, highlights important considerations for SEC-based complexome profiling methods. The SEC-SWATH-MS method was tested with the initial 120-minute gradient time, and a short-gradient SEC of only 30 minutes, the latter making this technique more high throughput (Bludau et al., 2020, 2023; Heusel et al., 2019). However, the meta-analysis showed that currently, longer gradient times are better in recovering lower-abundant proteins, although most co-fractionation MS methods identify the more abundant proteins (Figure 3A-B) (Skinnider & Foster, 2021). Furthermore, it was estimated that collecting more than 40 fractions did not significantly improve the identification of the known complexes from co-fractionation MS data, which was used as a benchmark to test the different methods (Figure 3C) (Skinnider & Foster, 2021).

The SEC-SWATH-MS methods should have collected a sufficient amount of fractions according to meta-analysis studies, as a total of 64 fractions and 81 fractions for the longer gradient time were collected. On the other hand, collecting data for more biological replicates, even more than five as considered in the meta-analysis, shows great improvement in the precision with which protein complexes are identified (Figure 3D) (Skinnider & Foster, 2021).

As it is desirable to use several biological replicates for which the same peptides ought to be covered, which is especially important in multiplexing studies where the peptides should be found in both reporter-channels to perform a relative quantification, carboxylate-modified (para)magnetic beads, like, for example, using SP3, should offer great improvements in the method. These beads have high peptide coverage, also allowing for lower sample input amounts. Furthermore, the SP3 method has high efficiency, allowing for clean-up and digestion to take place in a single tube (Havugimana et al., 2022; Hughes et al., 2014; Skinnider & Foster, 2021).

A further benefit of the SEC-SWATH-MS method, using the R-package CCprofiler for data analysis, is the possibility to extract also additional information from the SEC-data, whereby protein distribution over assembled and monomeric states, as well as over different complexes could be determined. This method approaches the data first from an assembled-mass level, whereby the distribution of assembled vs monomeric mass of proteins is estimated, then also at a protein-centric level, and finally at a complex-centric level (Figure 4).

The peptide data is filtered for only proteotypic peptides, which are the peptides that are most likely confidently quantified in MS studies, which reduces the possibility of false-positives. The peptide

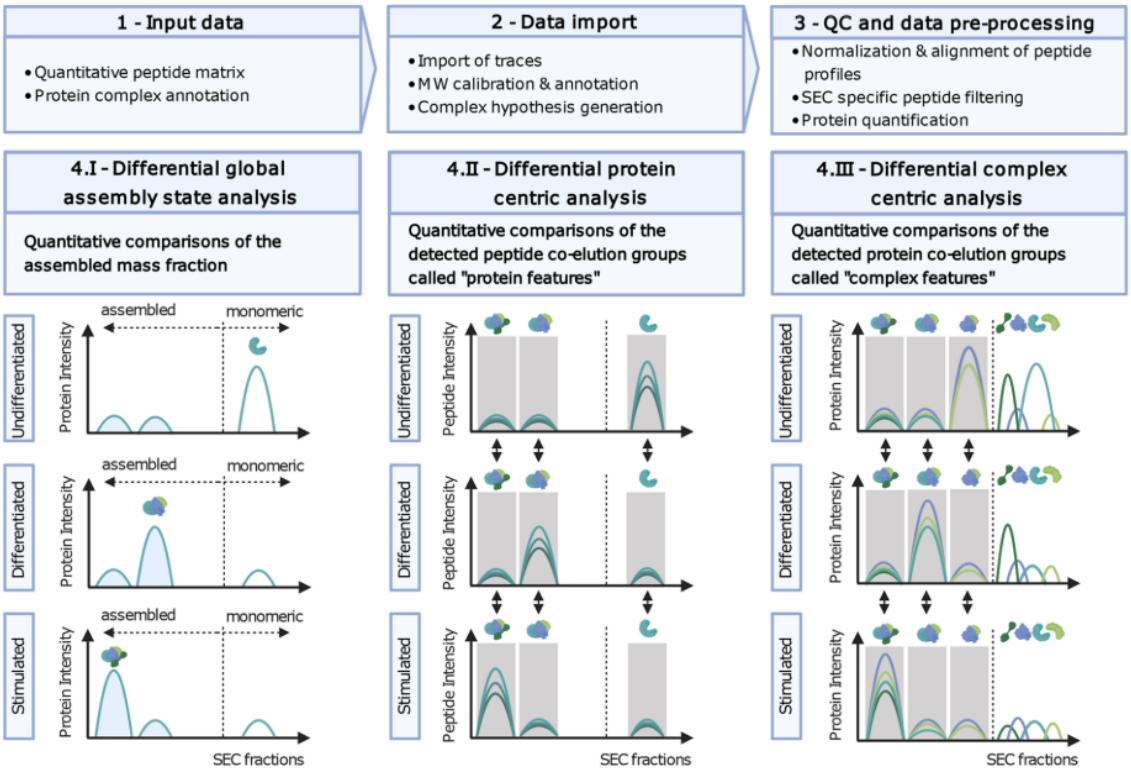

Figure 4: Figure from Bludau et al., 2023, showing the data analysis of CCprofiler. The top boxes contain information about the required input for each step. Panel 4.1 shows the principle of the global assembly state analysis. This reports the relative assembled fraction of a protein compared to its monomeric state. The assembled state is assigned when a protein elutes at molecular weights two times or more than its monomeric molecular weight. Panel 4.2 displays the protein features, whereby a different distribution of a protein over different complexes is assessed. In panel 4.3 the differential complex-centric analysis of CCprofiler is depicted. Different complex features are compared between different conditions.

precursor intensities are summed, to obtain quantitative data. To further reduce the chance of false-positives, only peptides occurring in a minimum of three consecutive fractions are taken into account, and they should have a high correlation with at least one other peptide originating from the same protein. It is assumed that peptide traces correspond to the SEC-profiles of the protein they belong to, as this should, in theory, be the case. The proteins are quantified by summing the top two peptide-traces, and normalization by the maximum. As also used for Blue-Native approaches, a calibration curve is used with standard proteins with a known mass, to annotate the fractions with the corresponding molecular mass, although this is more precise for more globular proteins. When proteins appear at a peak apex twice the size of the monomeric mass, the protein is considered to be in the assembled state. Thereby, it also allows us to elucidate if a protein is part of several different complexes. Upon further optimization of the method, it is also possible to compare the co-fractionation MS data of different conditions with each other, also on different levels.

This approach also takes into account if a difference in the distribution over different assembly states is due to a change in protein expression, or due to a change in the distribution. This is measured using the peptide signal intensities and taking the median of all peptides for a protein, allowing it also to perform statistical testing.

The complex-centric approach requires prior information about existing complexes, which could be derived from the CORUM database or the Bioplex network. This enables a decoy step ensuring that the complex-centric results could be FDR-controlled, which was not shown before. The downside is that it makes it less suitable to discover novel protein complexes with this approach. The protein traces are searched and tested against the protein co-elution features from the provided database. Also, this could be compared among the different conditions (Bludau et al., 2020, 2023; Heusel et al., 2019).

PREDICTION OF NOVEL PROTEIN COMPLEXES

In addition to the many benefits of CCprofiler, its major limitation is the unsuitability to identify novel complexes, as it requires prior knowledge of the complexes in the form of a user-provided database. This makes it less applicable for studies in which finding novel complexes, or several novel subunits of a complex are of interest, or if another organism has to be studied, which was not that extensively studied before and therefore lacks sufficient prior information.

Different machine-learning-based methods exist that can predict novel complexes from complexome profiling data. Some of the machine-learning approaches apply peak-centric approaches, whereby protein profiles are fitted to a model, e.g. a Gaussian model, from which the change in protein levels could be compared, as well as protein-protein interactions could be predicted using different distance metrics, like among others the Pearson correlation, using clustering strategies. Examples are PrInCE and ComplexFinder, both applicable to label-free and labeled data (Nolte & Langer, 2021; Stacey et al., 2017). To illustrate the benefits and limitations of these techniques, the more recent ComplexFinder will be described further, which is a Python-based computational approach, that can analyze mainly BN-PAGE and SEC data for different quantification strategies like LFQ, SILAC, or TMT. This method uses machine learning to predict protein-protein interactions from the fractionation data, whereby a peak-centric quantification approach is used to build the complexes and compare different conditions, from which a connectivity network can be reconstructed. Peaks are detected by finding local maxima in the protein feature profiles, whereby user-defined restrictions can be applied, allowing for peak-centric comparisons between samples. Each peak is fitted to several models, and after summing the best results, a signal profile is obtained. A peak alignment is used to correct for shifts within samples. For label-free samples, the Area Under the Curve (AUC) of the signal profile is used for quantification and identification of significantly different protein peaks, whereas for SILAC or TMT experiments, the full-width at half-maximum (FWHM) is used. Protein-protein interactions are predicted based on different distance calculations, like for example the commonly used Pearson or Spearman correlation among several others. Moreover, custom functions could be defined, with which external information could be added, for example, the subcellular localization. Generating additional decoy interactions, a classifier is trained to distinguish protein-protein interactions, after which for each protein-protein pair the probability of being an interaction is also calculated. A connectivity network is built, based on the identified protein pairs and the corresponding probabilities.

This software can predict interactions, and does not require the use of a reference database like in CCprofiler. However, it can integrate one, like the CORUM database to be used to predict protein interactions, or to provide additional information about the protein complexes, like subcellular localization. Important for peak-centric approaches like this, is the ability to describe the peak profile by a model.

If there are several overlapping peaks or the peak is at the noise level, thereby no longer resembling a Gaussian distribution, model fitting performs poorly, and this peak is filtered out based on its low fitting score (Nolte & Langer, 2021).

EPIC is another machine-learning platform, that uses clustering to form a protein connectivity network from the identified protein-protein interactions, from which complexes could be derived, but doesn't apply a peak-centric analysis as the previously mentioned methods do. It also supports additional information input, for example, from the CORUM database or Gene-Ontology databases, whereby this functional evidence might reduce false negatives. As for now, of

these toolkits, only ComplexFinder is suitable for the analysis of TMT data by a machine-learning-based approach (Hu et al., 2019; Nolte & Langer, 2021).

A downside of these methods, using clustering on the protein-protein interactions to decipher protein complexes, is that it is prone to noise-amplification. High-throughput strategies used in complexome profiling show a lot of 'noise', i.e. the biological and technical variations, causing errors in the 'edges' of the networks (the interactions) due to either false positive interactions or false negative interactions, leading to different outcomes, making these techniques less reproducible. An R-package is developed to test whether the clusters are more or less resistant to iterative rounds of network perturbations, whereby more stable clusters are more likely to be reproduced by different experiments, and should also more likely be biologically relevant (Stacey et al., 2021).

Another tool, PCprophet, was introduced to analyze co-fractionation MS data, suitable for different separation methods and quantification strategies, including TMT. This tool can detect novel protein complexes but does not depend on clustering analysis of protein-protein interactions into networks. As with CCprofiler, it uses a complex-centric approach, as the tool is trained with co-elution data of protein complexes, compared to the peak-centric and protein-protein-interaction-based approaches of e.g. PrInCE and ComplexFinder. In addition, statistical error models are performed to reduce false-positives. Furthermore, differential analysis between different conditions is used to find significant changes in the complexome. Besides its application of finding novel complexes, this method outperformed EPIC and CCprofiler in identifying known CORUM complexes. On the other hand, EPIC reported a higher amount of average subunits per complex, whereby the average reported for CCprofiler and PCprophet is closer to the amount in CORUM. Although not stated, this resemblance might be due to a bias towards CORUM-reported complex sizes, as this database is used as prior input in both CCprofiler and PCprophet, whereby the latter also uses training of the model by this database. Also, the 'node degree' distributions were compared among different networks derived from several tools and databases. The node degree indicates how many connections a single node has. In general, EPIC also shows more connections, thereby more closely resembling the STRING database-derived network instead of CORUM. PCprophet was also able to recover more closely connected proteins compared to EPIC. Several of the novel predicted complexes consist of a known complex from the CORUM database with an additional subunit that was not reported in the database before (Fossati et al., 2021).

As with CCprofiler, this method can search for differences between conditions on a protein- and a complex level, being able to differentiate between changes in protein abundance and assembly states on a protein level, and rewiring of complexes, including identification of changes in complex stoichiometries. To perform FDR-based controls, PCprophet still needs a protein complex or PPI database, like CORUM, as input. For this, GeneOntology (GO) terms are collected for each protein found in the complex and the similarities between each protein are then compared in a pair-wise way, resulting in an overall GO score for the complex. Of course, a high similarity between the GO scores for each protein pair in a protein complex is to expect for true positive complexes. This is also done for core complexes of the CORUM database, and by comparing the GO-score distributions for experimentally predicted novel complexes and the known database complexes, a GO score could then be selected that satisfies a certain FDR value and then used as a filtering criterium (Fossati et al., 2021).

At this moment there is a set of different machine learning-based tools that can be used to analyze co-fractionation MS data to compare different conditions, but also predict novel protein complexes, not reported in the literature before, in a high-throughput manner for the elaborate and complicated SEC-MS co-fractionation data. These tools all have to decide whether the predicted protein-protein interactions are positive interactions, which could be achieved by comparing it to information from available databases, to either integrate additional supporting evidence of the results from these user-provided databases containing publicly available data about reported complexes, to provide decoys or using GO scores for FDR-based error control as in PCprophet. Although ComplexFinder could be used without a database, the results are more accurate when additional information is integrated to validate the potential protein complexes. Furthermore, there still are differences in the output of all these tools, as was shown for example by the differences in average subunits per complex that could be found. Also, not all of the tools are suitable for TMT data. Therefore, the choice of data-analysis tool also depends on the method used, the availability of prior information, and what type of validations are appropriate. Of all these tools, PCprophet performed best in identifying known complexes from the CORUM database and was able to predict novel complexes in an error-controlled manner. The performance of PCprophet could be less when a less extensively reported organism in CORUM is used.

STUDYING PROTEOFORMS IN THE CONTEXT OF PROTEIN COMPLEXES

Complexome profiling offers great insight into biological systems, and how subtle changes in conditions might not only result in differential expression of proteins but importantly also enable catching changes in the rewiring of complexes. On the other hand, small changes that are less abundant, but have a major impact on the functionality, specificity,

subcellular localization, or assembly state of the complexes, might not directly be discovered by the techniques mentioned before. Post-translational modifications, PTMs, like for example phosphorylation, reversibly decorate the proteins and allow for fine-tuning cellular functions, whereby a protein's function depends on the specific modifications it has acquired at a specific moment. Finding the specific PTMs is difficult, especially when using SEC where whole lysates are mixed and peptides bearing PTMs become further diluted, therefore requiring specialized protocols to enrich for these peptides and to be able to trace this back to the correct proteins. Also, differences in PTM abundances should be attributed to either changes in the general protein expression or changes in the amount of PTM-modification events (Altelaar et al., 2013).

In the light of SEC-MS-based complexome profiling, different attempts have been made to accomplish this. In the earlier approach, a more general pipeline consisting of cell lysis, native SEC, and LC-MS/MS, was applied. A Superose 6 10/300 GL column was used, which accomplished the separation of complexes ranging from ~15kDa to approximately 1MDa into 40 fractions. Complexes were identified by searching for co-fractionation of known protein complexes, whereby a minimum amount of the proteins known to be part of the complex should show co-fractionation, or by using hierarchical clustering methods based on elution profile similarities and validated by the STRING database. The downside of clustering strategies is the use of a less objective cut-off, as the cut-off in the dendrogram on the number of clusters, has to be decided by the user. Additional information about possible protein-protein interactions, and possible PTMs, the data is integrated into the *Encyclopedia of Proteome Dynamics* database to infer protein properties to the proteomics data, like PTMs, alternative splicing events, subcellular localization, and turnover rates, and to test the likeability of potential protein-protein interactions. In addition, this method was mainly applicable to the study of known complexes, as database information was used as validation for the results. As some proteins can be part of several complexes, they elute in several fractions during SEC separation and are also found in the network in different clusters. Phosphorylation was studied by looking at the specific phosphorylated peptides that were detected by MS and tracing back to which peak fractions it originated and inferring this information into the clustering network. Some proteoforms of a protein were exclusively found to belong to a specific protein complex subset, as the phosphorylated peptide was not found in all elution fractions of the protein, but only in a specific one, whereby the different elution fractions clustered together to form different protein complexes. This illustrates the function of phosphorylation to regulate protein-protein interactions and determine which protein complex should be assembled (Kirkwood et al., 2013). This strategy was successful in deciphering some phosphorylation patterns but didn't use any enrichment strategy, making it highly likely that a lot of PTMs are lost in the analysis. Furthermore, the lysate contains a mix of all different PTM states of the protein, making it difficult to assign a specific biological function to a phosphorylation site, as it is not known in which condition a certain PTM is applicable. Furthermore, it would be interesting to also add the information from the *Encyclopedia of Proteome Dynamics* they used in the context of PTMs, by looking at the subcellular localizations of the proteins in the cluster containing a specific PTM and the cluster without it, and see if there might be correlations in PTM-profiles and the subcellular localization. This could also be done by looking at the turnover rates.

The COrrelation-based functional ProteoForm (COPF) assessment tool was developed to detect and assign proteoforms more systematically in SEC-MS experiments, mainly for DIA-MS experiments, as these provide additional sensitivity and coverage of distinct proteoforms (Figure 5). This method was able to determine cell cycle-, or tissue-specific proteoforms. If a protein has no differences in proteoforms in the tested conditions, all of the peptides derived from this protein should show similar quantitative profiles, whereas proteins that show differential proteoform expression, show a set of different quantitative profiles and distinct proteoform groups could be assigned. By taking these considerations into account, a strategy is developed to calculate a proteoform score that informs whether a protein has differentially behaving proteoforms. For this, all peptides belonging to a protein are hierarchically clustered and divided into two separate clusters, containing a minimum of two peptides in each cluster, whereby it is assumed that peptides belonging to a specific proteoform cluster together. Then the Pearson correlation scores are calculated, both for all of the peptides and for all peptides within a cluster, whereby for the latter the lowest correlation value is selected as the within-cluster correlation score. The proteoform score is calculated by subtracting the within-cluster score from the overall cluster score, whereby a higher score indicates differentially behaving proteoforms, and also corresponding p-values could be estimated. This tool can also be integrated into the CCprofiler framework, enabling the detection of assembly-specific proteoforms. A great benefit of this tool is the possibility to compare different conditions with each other, provide statistical analysis, FDR estimation for proteoform detection, and decipher assembly-, or condition-specific proteoform groups. As COPF makes use of the inherent variation of the data to test for different proteoforms of a protein, it performs less for small changes, like proteoforms that differ by only one peptide. In a benchmarking study, phosphosites could be identified using this approach. Based on information from the literature, known phosphosites could be selected for a targeted re-analysis of the SEC-SWATH data, although not all of the possible phosphosites could be reliably detected. The ones that

were detected, fell into the same proteoform group as determined by COPF. Although this approach can detect some phospho-specific proteoform groups, without the use of phosphatase inhibitors and phospho-enrichment, the protocol could still be upgraded to find even more phosphosites. To resolve more than two proteoform groups, different clustering methods should be chosen by the user, which are already available in COPF (Bludau et al., 2021).

The progress in data-analysis strategies allows for the subtraction of as much data as possible from each dataset, and with the use of more sensitive MS methods, like DIA, proteins are more reliably detected, as well as the discovery of less abundant proteins, complexes, or different proteoforms. Furthermore, the amount of data that is gathered into publicly available databases allows for the implementation of additional evidence for protein-protein interactions, and integration of additional information like subcellular localization.

The integration of COPF into the CCprofiler tool enables the detection of different proteoforms, for single conditions or multiple conditions tested, and assign them to assembly-specific proteoform groups. This method also has the benefit of providing FDR estimations.

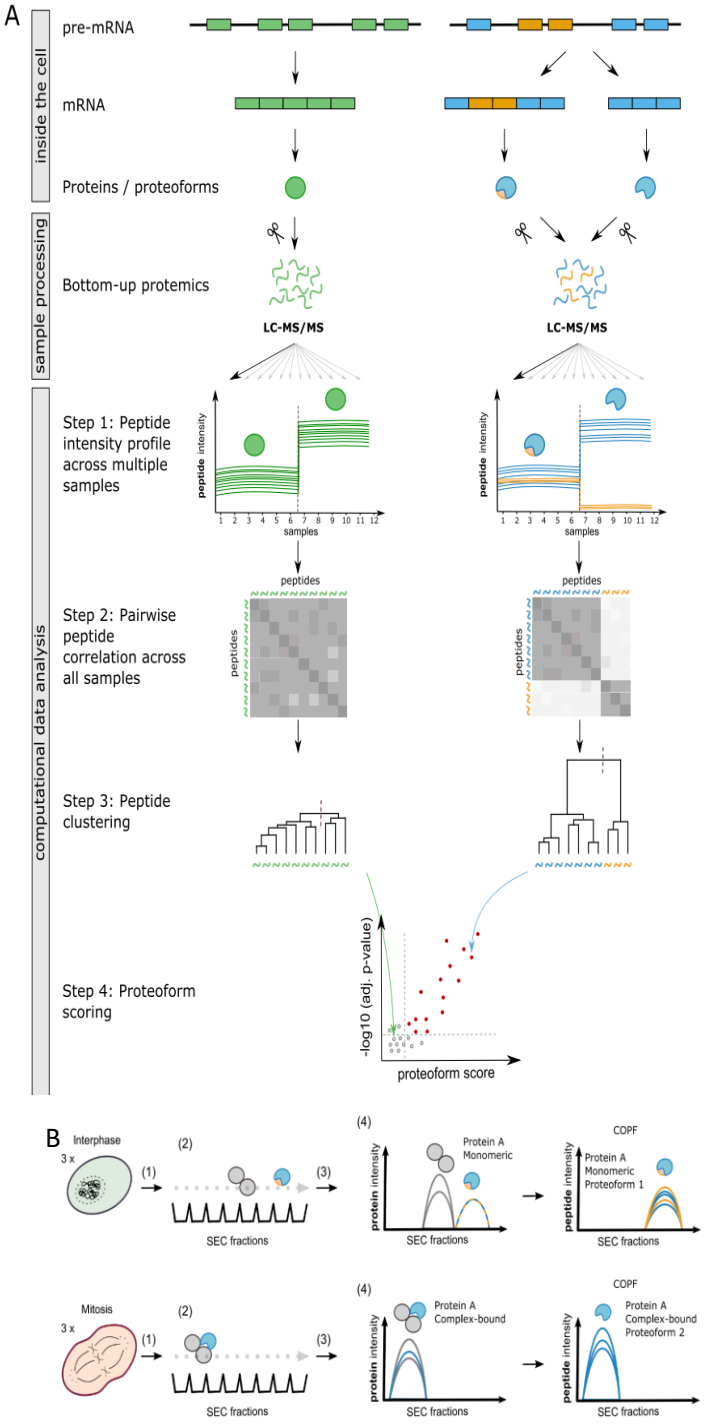

Figure 5: Figure from Bludau et al., 2021. COPF is based on the concept that for proteins with a single proteoform, the peptides show similar quantitative profiles. A) Schematic overview of COPF. When several proteoforms exist, peptides belonging to a specific proteoform should have a distinct quantitative profile, shown in orange (Step 1). This example shows a set of 12 samples, across two different conditions. Peptides are correlated in a pair-wise way for each protein (Step 2). Based on the correlation distance, peptides are hierarchically clustered, and separated into two groups (Step 3). Peptides derived from the same proteoform should cluster together, and show higher correlation scores within the cluster. A proteofrom score is calculated to compare the overall correlation score, with the within-cluster correlation score. Proteins containing proteoforms have higher proteofrom scores. P-values are calculated (Step 4). B) Using a SEC-SWATH-MS analysis workflow, and for example CCprofiler, condition- or assembly-specific proteoforms can be elucidated with this technique.

DISCUSSION AND CONCLUSION

SEC is very suitable for studying protein complexes in a high-throughput way and with CCprofiler even in an FDR-controlled way. The native purification with SEC has the added benefit of causing less likely perturbations from tagging, e.g. disrupting the complex by preventing some interactions, disturbing its function, or changing its cellular location, which might occur in targeted approaches where a tag or enzyme is added to the proteins of interest. However, limitations from the experimental set-up with SEC still exist. Most common SEC approaches use a bit more sample input compared to for example BN-PAGE, requiring around a milligram of material.

On one hand, SEC performs highly in its identification of native soluble complexes, but a lot of membrane proteins and complexes are undetected and less likely to be identified. This is due to mild detergents being used to keep the soluble protein complexes in native conditions, while this is not optimal for most membrane proteins, which require more hydrophobic environments using detergents not compatible with MS analysis. Specialized methods are developed to solubilize membrane proteins in native conditions, like nanodiscs or peptidiscs. After cell lysis using mild detergents, the membrane fraction is reconstituted immediately in a membrane-mimicking environment, like peptidiscs, that wrap around the membrane proteins. This reduces the chance of membrane protein dissociation and disassembly, by keeping the membrane complexes in their native environment. These membrane mimetics are soluble, and for the following SEC-fractionation and MS steps, no detergent has to be used. This has already been shown to give nice results for membrane complexes in *E. coli* (Luke Carlson et al., 2019; Salas et al., 2020; Skinnider & Foster, 2021).

During lysis in the buffer volume proteins are diluted, which might disrupt weaker complexes. This problem occurs for all strategies, but in SEC further dilution occurs in the chromatography steps where the samples are mixed with the native buffer used. Furthermore, protein complexes that are more thermodynamically labile, having high off-rates, might disintegrate more during the SEC procedure, limiting their detection (Heusel et al., 2019). Besides missing information about more transient interactions, which are still highly relevant for biological functions, less-abundant protein complexes are also more easily missed. In addition, dilution increases the spreading of proteins over the column thereby reducing the resolution that could be achieved. Furthermore, self-oligomers can be formed due to protein aggregation, which might influence the calculated assembled/monomeric distribution in CCprofiler, whereby these artificial aggregations also add up to the assembled state mass while they are not biologically relevant (Burgess, 2018; Cabrera-Orefice et al., 2022; Heusel et al., 2019; Iacobucci et al., 2021; Skinnider & Foster, 2021).

Concerning the preservation of more transient interactions, crosslinking MS (XL-MS) has been combined with several established complexome profiling techniques, like BN-PAGE, Proximity Labeling (PL), and SEC. XL-MS stabilizes weaker/more transient interactions, although this might also capture non-relevant interactions of proteins that are nearby by chance (Hevler et al., 2021; Larance et al., 2016; Liu et al., 2020; Wang et al., 2022). However, in-gel cross-linking showed fewer over-length cross-links, most often caused by protein aggregation, and unspecific cross-links compared to in-solution cross-linking. It has the additional benefits of having no necessary optimization steps to determine optimal crosslinker and protein concentrations, thereby using less sample, and being able to determine conformation-specific cross-links. The latter is important for studying co-occurring protein complexes that share subunits, whereby in-solution cross-linking shows a mixture of all the different assembly states, whereas in-gel cross-linking is more targeted, enabling the elucidation of distinct assembly states of protein subunits into different complexes with the cross-links and their related distance restraints being provided for each separately. This provides assembly- and conformation-specific information (Hevler et al., 2021). The benefit of the CCprofiler algorithm is the ability to also distinguish different assembly states. Cross-linking for in-solution samples, which will be fractionated by SEC-SWATH-MS, combined by analysis with CCprofiler would be an interesting approach to obtain high-throughput, FDR-controlled data of also more transient and less abundant complexes, whereby also the different assembly states can be elucidated. Another benefit of crosslinking is the possibility to also aid in studying membrane proteins (Larance et al., 2016).

Moreover, it is very important to keep in mind that it is challenging to have a standard to which the different complexome profiling techniques could be benchmarked and which can be used to determine how close to real native conditions our experimental outcomes are, as it is almost impossible to define the 'ground-truth' condition. The complexes that are formed inside the cell depend on the cell signaling events that inform the cell about its environment to which it has to adjust. This means that the complexome state changes over time, as cells undergo different stimuli and cell cycle phases, whereby complexes need to adjust to meet the requirements at a specific time. Also, the type of cells used could differ from other cell types in their complexome state, moreover, they could differ from what is happening inside the context of organs, or even the whole organism. Therefore, it is important to choose the correct system to study your topic of interest, and to select a proper control to be able to test the different conditions, whereby ideally only differences are found due to the change of condition and not due to other artifacts, like different cell types, or culturing media. Furthermore, if several cells are lysed and this mixture is analyzed, the results resemble a convolution of all

of the different cells and their complexome states. Also, protein isoforms are diluted in the context of the general most common protein form. In addition, upon lysis of whole cells, subcellular localization information is lost, while complexes might show preference for specific locations inside the cells, or might have different functions or stoichiometries depending on their location.

To also retrieve localization information and more dynamic interactions, several approaches are developed. This is more straightforward when using targeted approaches, as these methods already select specific complexes, compared to SEC, which is mostly used to look at the whole proteome and complexome at once. Proximity-labeling (PL) approaches can be used when only a subproteome, like the nucleus or ribosome, is of interest, as these target a specific complex and label proteins within a specific range around it, capturing proteins within a radius of about 10-20nm. When using APEX labeling, these approaches can achieve a high temporal resolution, with labeling times of approximately one minute, capturing also highly dynamic interactions (Dionne & Gingras, 2022; Ke et al., 2021; Liu et al., 2020). Combining proximity labeling with cross-linking also provides great opportunities, as proximity labeling allows for subproteome-level targeting, whereby XL-MS delivers information on a smaller scale, about 1nm, to distinguish direct and indirect binders from PL studies, but are also able to target proteins that are in close contact with each other and are also part of the complex, but might not be targeted by the bait protein of the PL assay directly, resulting in a more comprehensive protein network analysis. Furthermore, due to the applied XL-MS, the approach is less sensitive to nonspecific binding, which is a common caveat of targeted approaches (Liu et al., 2020). Fascinating is the progress in data-analysis strategies that are developed, whereby subcellular localization information could be implemented for SEC data when using custom functions in ComplexFinder (Nolte & Langer, 2021). For further research, it would be interesting to implement these types of calculations in other data-analysis software, like CCprofiler to obtain an even more extensive overview of the whole complexome state at higher spatial resolution. These methods should be optimized further to look more specifically at complexome differences at subcellular locations for SEC data.

To even retrieve a more complete picture of the complexome state of the cell upon different conditions, methods to also define assembly-specific proteoforms are available, like COPF. However, this does not apply to the detection of novel protein complexes, as it is dependent on CCprofiler, which has to use a provided database. Furthermore, it can detect changes in proteoform assemblies in changing conditions, and GO-terms could be added from existing databases to also see differences in subcellular localization. The latter is then only possible for cells, or organisms and conditions that have been studied before. Methods like proximity assays look at subcellular localization and provide higher spatial resolution to define the localization.
Furthermore, improvements in the protocol could be made to cover more PTMs, like phosphorylation, by applying other fragmentation strategies, instead of CID or HCD which break weaker bonds first, like ETD or ECD, which are better at preserving labile bonds. ETD has already been shown to be compatible with DIA-MS and should provide higher coverage of the less abundant proteoforms (Doll & Burlingame, 2015; Schmidlin & Altelaar, 2020) .

A way to improve the detection of phosphopeptides is by using AP-MS, which increases the sensitivity to detect phosphopeptides by enriching them. However, a pitfall of these targeted strategies like AP-MS, IP-MS, and proximity labeling is that they show a convoluted image of the proteoforms. All of the different proteoform-specific complexes might be simultaneously purified since the complexes that are purified depend on binding to the bait protein or antibody, whereby some proteins might show less affinity in either the presence or absence of the modification, but other proteins of the complex might still be able to bind. As a result, it is showing a convolution of all subcomplexes that were concurrently purified, and also missing the information about possible modifications on other proteins that are part of the complex. Thereby, the specific effect of a set of modifications on protein complex formation might be lost and the specific proteoform-specific subcomplexes might not be correctly elucidated. Furthermore, in AP-MS, specific antibodies targeting the specific modifications should be used, but these might not be available for all the specific sites a protein has.

In correlation-profiling-based approaches, the co-elution or migration profiles show differences for specific proteoform groups, as already shown by Bludau et al 2021, which has been implemented in COPF. These types of methods can show distinct assembly-specific proteoforms, whereby a specific modification might only be present in the elution profile of one subcomplex. Due to the separation in specific subcomplexes, it is possible to determine on which modification the formation of a specific subcomplex might depend.
Phospho-DIFFRAC is also another approach that can be used to study phosphorylation-dependent protein complex assemblies, which compares elution profiles of non-specific phosphatase-treated and phosphatase-inhibitor treated conditions (Floyd et al., 2021). This latter technique does not preserve specific biologically relevant phosphorylation patterns and the corresponding phospho-specific subcomplexes. This is due to the non-specific treatments used, whereby complexes dependent on the specific phosphorylation of only some of the phospho-sites might be lost.

Combining a targeted strategy with correlation profiling should allow for a more sensitive approach to detect proteoforms, while also preserving the association between the modifications and complex-formation. This was already explored for YAP1 (Figure 6). In this study, the different YAP1-complexes were first purified by AP-MS, in the presence or absence of phosphatase inhibitors. A downside of this method is that a phosphatase inhibitor is not acting on specific phospho-sites, and if more subcomplexes exist, each depending on the phosphorylation of a different phospho-site, these might not be completely and convincingly elucidated by AP-MS. However, via a clustering approach, protein-protein interactors are indicated that show differential affinity to YAP1 based on the serine/threonine or tyrosine phosphatase inhibitor that was used. Therefore, the purified complexes are then separated by BN-PAGE and analyzed by MS to elucidate the composition of the complexes in the different YAP1 phosphorylation states. BN-PAGE is better at resolving low-abundant co-migrating complexes, like those of phospho-peptides, than SEC is. Nine different subcomplexes, each with a specific phosphorylation pattern, were discovered by hierarchical clustering of the BN-PAGE migration profiles. Each of the subcomplexes found could be linked to different target functions of YAP1. To test for different phospho-sites, specific phospho-mutants have been created, whereby AP-MS and co-migration data were integrated, showing that several phospho-sites regulate subcomplex integrity, whereby in almost all of these cases the phospho-site is necessary for binding, as confirmed by AP-MS. Several Knock-Out (KO) cells were made, each lacking a key regulator of the signaling pathway of which YAP1 is part. As these KO strategies are less compatible with ectopically expressing YAP1, IP-MS was used instead of AP-MS. This links the role of several pathway regulators to the phosphorylation patterns of YAP1 and its effect on complex formation. Still, some caveats exist within this protocol, as large amounts of sample input are needed (Uliana et al., 2023).

The use of AP-MS instead of IP-MS might provide some limitations for this technique. Suitable antibodies are not always available, and antibodies might not always interact with different proteoforms, or all interaction partners due to the overlap of the interacting domain with the antibody epitope, thereby adding a possible limit to co-purify some proteo-form specific complexes. For AP-MS, the bait protein can accumulate different modifications inside the cell in which it is expressed, while still being able to bind the different subset of interaction partners, and whereby the tag is still able to bind to the beads used for purification. Thereby, this method increases the chance of co-purification of all different proteoform-specific complex assemblies.

As a major limit, for all strategies applied, there remains a risk of complexes that may disintegrate during the purification, resulting in partially resolved complexes, but also the risk of co-elution due to similarity in biophysical properties on which the separation is based, or the risk of protein aggregation. Therefore, it is important to see if the observed complexes represent in vivo conditions, by using proper controls, or by using a likelihood prediction, which CCprofiler does. That is a great benefit of approaches like CCprofiler, but the experimental results could also be compared with prior information about the complex manually if this information is available, although thereby not providing statistical evidence. To be able to provide this for novel complex prediction, PCprophet was developed, allowing for error control based on GO-terms. However, this could perform less when it is applied to organisms for which no complete GO database exists. Additional testing for possible protein-protein interactions could nowadays be done by using AI methods, that can predict protein complexes by machine learning, additionally providing information about the structure as a great benefit (Humphreys et al., 2021; Shor & Schneidman-Duhovny, 2024). This could be integrated as an additional validation tool in programs like CCprofiler or PCprophet, in cases where sufficient databases do not exist, and might also add structural information to the obtained complexes. Of course, it still requires an awareness of possible false positives or negatives that might then be integrated. In addition, prediction-based methods, like PCprophet and ComplexFinder, might be better at describing the complexome state more accurately for specific conditions, whereas the database provided for methods like CCprofiler only contains information about certain conditions that have been studied before, and might not be accurate enough to explain the complexome in another cell type or condition. A more specific database, in concordance with the sample used and the condition that is tested, could also improve the accuracy of the results.

Moreover, assembly intermediates might be confused with partially disintegrated complexes, while they could provide additional information about the biological function of a complex. BN-PAGE is better at finding assembly intermediate states, compared to SEC-based methods, due to its high-resolution separation. With CCprofiler some intermediates might be found upon manual inspection, which is labor-intensive. With ComplexFinder, it is currently not possible to find the intermediates, as the analysis is performed on the protein level, whereby each complex found is treated as a unique complex. For future research, it might also be interesting to implement strategies to identify the intermediate complexes in a more automated method, although this increases the complexity of the data analysis and requires additional statistical methods to determine whether a complex might be an intermediate complex or a partially disassembled complex. This is yet also information that is missing from AI-based approaches, as these predict also the protein structure, and the possible complexes, but do not show intermediate states, which might also have important biological functions, or could be of interest as drug targets.

To conclude, it is worth mentioning that we should keep in mind what kind of complexes need to be studied and in which context we want to study it, whether we expect to identify novel subunits or completely novel complexes, if the interactions are transient or highly dynamic, if the protein is highly abundant, or if it has a preferred subcellular localization. This is important in determining which is the optimal strategy to apply. For example, in the elucidation of proteoform-specific complex assemblies, SEC provides a nice way to separate the different subcomplexes and preserve the proteoform-specific information, although it might not be sensitive enough. Therefore, it has been combined with targeted approaches like AP-MS or IP-MS. In general, SEC is an ideal method to study native protein complexes, which can be adjusted to a high-throughput manner, and is more cost-effective compared to targeted approaches, as no expensive antibodies need to be bought, or several proteins need to be tagged. Furthermore, with the FDR-controlled analysis from CCprofiler, it is ideal to compare different conditions and gain an impression of the overall complexome state but also to identify different subcomplexes, and protein distributions over the monomeric state, or different assembly states.

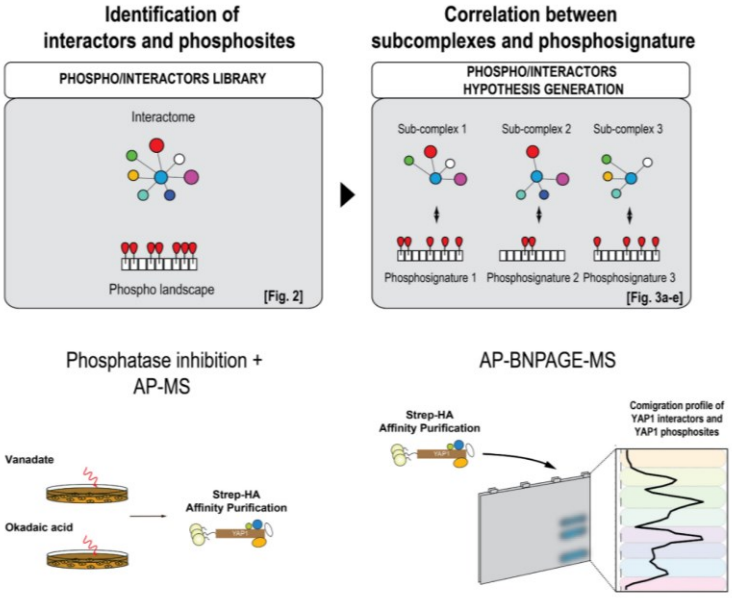

Figure 6: From Uiliana et al., 2023. Schematic drawing of the strategy used to study the phosphorylation-dependent complex organization of YAP1, using a combination of AP-MS and correlation profiling with BN-PAGE. The interactors and phosphosites are identified by AP-MS. These purified complexes are then separated by BN-PAGE-MS to decipher distinct assemblies.

REFERENCES

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. In *Nature* (Vol. 537, Issue 7620, pp. 347–355). Nature Publishing Group. https://doi.org/10.1038/nature19949

Altelaar, A. F. M., Munoz, J., & Heck, A. J. R. (2013). Next-generation proteomics: Towards an integrative view of proteome dynamics. In *Nature Reviews Genetics* (Vol. 14, Issue 1, pp. 35–48). https://doi.org/10.1038/nrg3356

Bludau, I., Frank, M., Dörig, C., Cai, Y., Heusel, M., Rosenberger, G., Picotti, P., Collins, B. C., Röst, H., & Aebersold, R. (2021). Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24030-x

Bludau, I., Heusel, M., Frank, M., Rosenberger, G., Hafen, R., Banaei-Esfahani, A., van Drogen, A., Collins, B. C., Gstaiger, M., & Aebersold, R. (2020). Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes. *Nature Protocols*, *15*(8), 2341–2386. https://doi.org/10.1038/s41596-020-0332-6

Bludau, I., Nicod, C., Martelli, C., Xue, P., Heusel, M., Fossati, A., Uliana, F., Frommelt, F., Aebersold, R., & Collins, B. C. (2023). Rapid Profiling of Protein Complex Reorganization in Perturbed Systems. *Journal of Proteome Research*, *22*(5), 1520–1536. https://doi.org/10.1021/acs.jproteome.3c00125

Budayeva, H. G., & Cristea, I. M. (2014). A mass spectrometry view of stable and transient protein interactions. *Advances in Experimental Medicine and Biology*, *806*, 263–282. https://doi.org/10.1007/978-3-319-06068-2_11

Burgess, R. R. (2018). A brief practical review of size exclusion chromatography: Rules of thumb, limitations, and troubleshooting. In *Protein Expression and Purification* (Vol. 150, pp. 81–85). Academic Press Inc. https://doi.org/10.1016/j.pep.2018.05.007

Cabrera-Orefice, A., Potter, A., Evers, F., Hevler, J. F., & Guerrero-Castillo, S. (2022). Complexome Profiling—Exploring Mitochondrial Protein Complexes in Health and Disease. In *Frontiers in Cell and Developmental Biology* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fcell.2021.796128

Clucas, J., & Meier, P. (2023). Roles of RIPK1 as a stress sentinel coordinating cell survival and immunogenic cell death. In *Nature Reviews Molecular Cell Biology* (Vol. 24, Issue 11, pp. 835–852). Nature Research. https://doi.org/10.1038/s41580-023-00623-w

Dionne, U., & Gingras, A. C. (2022). Proximity-Dependent Biotinylation Approaches to Explore the Dynamic Compartmentalized Proteome. In *Frontiers in Molecular Biosciences* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fmolb.2022.852911

Doll, S., & Burlingame, A. L. (2015). Mass spectrometry-based detection and assignment of protein posttranslational modifications. In *ACS Chemical Biology* (Vol. 10, Issue 1, pp. 63–71). American Chemical Society. https://doi.org/10.1021/cb500904b

Floyd, B. M., Drew, K., & Marcotte, E. M. (2021). Systematic Identification of Protein Phosphorylation-Mediated Interactions. *Journal of Proteome Research*, *20*(2), 1359–1370. https://doi.org/10.1021/acs.jproteome.0c00750

Fossati, A., Li, C., Uliana, F., Wendt, F., Frommelt, F., Sykacek, P., Heusel, M., Hallal, M., Bludau, I., Capraz, T., Xue, P., Song, J., Wollscheid, B., Purcell, A. W., Gstaiger, M., & Aebersold, R. (2021). PCprophet: a framework for protein complex prediction and differential analysis using proteomic data. *Nature Methods*, *18*(5), 520–527. https://doi.org/10.1038/s41592-021-01107-5

Gnanasekaran, P. , P. H. R. (2023). Affinity Purification-Mass Spectroscopy (AP-MS) and Co-Immunoprecipitation (Co-IP) Technique to Study Protein–Protein Interactions. In S. Mukhtar (Ed.), *Protein-Protein Interactions. Methods in Molecular Biology* (Vol. 2690, pp. 81–85). Humana.

Guerrero-Castillo, S., Krisp, C., Küchler, K., Arnold, S., Schlüter, H., & Gersting, S. W. (2021). Multiplexed complexome profiling using tandem mass tags. *Biochimica et Biophysica Acta - Bioenergetics*, *1862*(9). https://doi.org/10.1016/j.bbabio.2021.148448

Havugimana, P. C., Goel, R. K., Phanse, S., Youssef, A., Padhorny, D., Kotelnikov, S., Kozakov, D., & Emili, A. (2022). Scalable multiplex co-fractionation/mass spectrometry platform for accelerated protein interactome discovery. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-31809-z

Heusel, M., Bludau, I., Rosenberger, G., Hafen, R., Frank, M., Banaei-Esfahani, A., van Drogen, A., Collins, B. C., Gstaiger, M., & Aebersold, R. (2019). Complex-centric proteome profiling by SEC - SWATH - MS . *Molecular Systems Biology*, *15*(1). https://doi.org/10.15252/msb.20188438

Hevler, J. F., Lukassen, M. V, Cabrera-Orefice, A., Arnold, S., Pronker, M. F., Franc, V., & Heck, A. J. R. (2021). Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry. *The EMBO Journal*, *40*(4). https://doi.org/10.15252/embj.2020106174

Hu, L. Z. M., Goebels, F., Tan, J. H., Wolf, E., Kuzmanov, U., Wan, C., Phanse, S., Xu, C., Schertzberg, M., Fraser, A. G., Bader, G. D., & Emili, A. (2019). EPIC: software toolkit for elution profile-based inference of protein complexes. *Nature Methods*, *16*(8), 737–742. https://doi.org/10.1038/s41592-019-0461-4

Hughes, C. S., Foehr, S., Garfield, D. A., Furlong, E. E., Steinmetz, L. M., & Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular Systems Biology*, *10*(10). https://doi.org/10.15252/msb.20145625

Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T. J., Banjade, S., Bagde, S. R., Stancheva, V. G., Li, X. H., Liu, K., Zheng, Z., Barrero, D. J., Roy, U., Kuper, J., Fernández, I. S., Szakal, B., … Baker, D. (2021). Computed structures of core eukaryotic protein complexes. *Science*, *374*(6573). https://doi.org/10.1126/science.abm4805

Iacobucci, I., Monaco, V., Cozzolino, F., & Monti, M. (2021). From classical to new generation approaches: An excursus of -omics methods for investigation of protein-protein interaction networks. In *Journal of Proteomics* (Vol. 230). Elsevier B.V. https://doi.org/10.1016/j.jprot.2020.103990

Ke, M., Yuan, X., He, A., Yu, P., Chen, W., Shi, Y., Hunter, T., Zou, P., & Tian, R. (2021). Spatiotemporal profiling of cytosolic signaling complexes in living cells by selective proximity proteomics. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-020-20367-x

Kirkwood, K. J., Ahmad, Y., Larance, M., & Lamond, A. I. (2013). Characterization of native protein complexes and protein isoform variation using sizefractionation- based quantitative proteomics. *Molecular and Cellular Proteomics*, *12*(12), 3851–3873. https://doi.org/10.1074/mcp.M113.032367

Kong, Q., Ke, M., Weng, Y., Qin, Y., He, A., Li, P., Cai, Z., & Tian, R. (2022). Dynamic Phosphotyrosine-Dependent Signaling Profiling in Living Cells by Two-Dimensional Proximity Proteomics. *Journal of Proteome Research*, *21*(11), 2727–2735. https://doi.org/10.1021/acs.jproteome.2c00418

Krasny, L., & Huang, P. H. (2021). Data-independent acquisition mass spectrometry (DIA-MS) for proteomic applications in oncology. In *Molecular Omics* (Vol. 17, Issue 1, pp. 29–42). Royal Society of Chemistry. https://doi.org/10.1039/d0mo00072h

Larance, M., Kirkwood, K. J., Tinti, M., Murillo, A. B., Ferguson, M. A. J., & Lamond, A. I. (2016). Global membrane protein interactome analysis using in vivo crosslinking and mass spectrometry-based protein correlation profiling. *Molecular and Cellular Proteomics*, *15*(7), 2476–2490. https://doi.org/10.1074/mcp.O115.055467

Leland H. Hartwell, J. J. H. S. L. & A. W. M. (1999). From molecular to modular cell biology. *Nature*, *402*(6761), C47–C52.

Liu, C. H., Chien, M. J., Chang, Y. C., Cheng, Y. H., Li, F. A., & Mou, K. Y. (2020). Combining Proximity Labeling and Cross-Linking Mass Spectrometry for Proteomic Dissection of Nuclear Envelope Interactome. *Journal of Proteome Research*, *19*(3), 1109–1118. https://doi.org/10.1021/acs.jproteome.9b00609

Low, T. Y., Syafruddin, S. E., Mohtar, M. A., Vellaichamy, A., A Rahman, N. S., Pung, Y. F., & Tan, C. S. H. (2021). Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. In *Cellular and Molecular Life Sciences* (Vol. 78, Issue 13, pp. 5325–5339). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00018-021-03856-0

Ludvigsen, M., & Honoré, B. (2018). Transcriptomics and Proteomics: Integration? In *Encyclopedia of Life Sciences* (pp. 1–7). Wiley. https://doi.org/10.1002/9780470015902.a0006188.pub2

Luke Carlson, M., Greg Stacey, R., William Young, J., Singh Wason, I., Zhao, Z., Rattray, D. G., Scott, N., Kerr, C. H., Babu, M., Foster, L. J., & Duong Van Hoa, F. (2019). *Profiling the Escherichia coli membrane protein interactome captured in Peptidisc libraries*. https://doi.org/10.7554/eLife.46615.001

Mann, M., Kulak, N. A., Nagaraj, N., & Cox, J. (2013). The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. In *Molecular Cell* (Vol. 49, Issue 4, pp. 583–590). https://doi.org/10.1016/j.molcel.2013.01.029

Morris, J. H., Knudsen, G. M., Verschueren, E., Johnson, J. R., Cimermancic, P., Greninger, A. L., & Pico, A. R. (2014). Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nature Protocols*, *9*(11), 2539–2554. https://doi.org/10.1038/nprot.2014.164

Muller, C. S., Bildl, W., Haupt, A., Ellenrieder, L., Becker, T., Hunte, C., Fakler, B., & Schulte, U. (2016). Cryo-slicing blue native-mass spectrometry (csBN-MS), a Novel technology for high resolution complexome profiling. *Molecular and Cellular Proteomics*, *15*(2), 669–681. https://doi.org/10.1074/mcp.M115.054080

Müller, C. S., Bildl, W., Klugbauer, N., Haupt, A., Fakler, B., & Schulte, U. (2019). High-resolution complexome profiling by cryoslicing bn-ms analysis. *Journal of Visualized Experiments*, *2019*(152). https://doi.org/10.3791/60096

Nolte, H., & Langer, T. (2021). ComplexFinder: A software package for the analysis of native protein complex fractionation experiments. *Biochimica et Biophysica Acta - Bioenergetics*, *1862*(8). https://doi.org/10.1016/j.bbabio.2021.148444

Páleníková, P., Harbour, M. E., Ding, S., Fearnley, I. M., Van Haute, L., Rorbach, J., Scavetta, R., Minczuk, M., & Rebelo-Guiomar, P. (2021). Quantitative density gradient analysis by mass spectrometry (qDGMS) and complexome profiling analysis (ComPrAn) R package for the study of macromolecular complexes. *Biochimica et Biophysica Acta - Bioenergetics*, *1862*(6). https://doi.org/10.1016/j.bbabio.2021.148399

Salas, D., Stacey, R. G., Akinlaja, M., & Foster, L. J. (2020). Next-generation interactomics: Considerations for the use of co-elution to measure protein interaction networks. In *Molecular and Cellular Proteomics* (Vol. 19, Issue 1, pp. 1–10). American Society for Biochemistry and Molecular Biology Inc. https://doi.org/10.1074/mcp.R119.001803

Schmidlin, T., & Altelaar, M. (2020). Effects of electron-transfer/higher-energy collisional dissociation (EThcD) on phosphopeptide analysis by data-independent acquisition. *International Journal of Mass Spectrometry*, *452*. https://doi.org/10.1016/j.ijms.2020.116336

Shor, B., & Schneidman-Duhovny, D. (2024). CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. *Nature Methods*. https://doi.org/10.1038/s41592-024-02174-0

Skinnider, M. A., & Foster, L. J. (2021). Meta-analysis defines principles for the design and analysis of co-fractionation mass spectrometry experiments. *Nature Methods*, *18*(7), 806–815. https://doi.org/10.1038/s41592-021-01194-4

Smits, A. H., & Vermeulen, M. (2016). Characterizing Protein–Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. In *Trends in Biotechnology* (Vol. 34, Issue 10, pp. 825–834). Elsevier Ltd. https://doi.org/10.1016/j.tibtech.2016.02.014

Stacey, R. G., Skinnider, M. A., & Foster, L. J. (2021). On the Robustness of Graph-Based Clustering to Random Network Alterations. *Molecular and Cellular Proteomics*, *20*. https://doi.org/10.1074/MCP.RA120.002275

Stacey, R. G., Skinnider, M. A., Scott, N. E., & Foster, L. J. (2017). A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics*, *18*(1). https://doi.org/10.1186/s12859-017-1865-8

Strecker, V., Wumaier, Z., Wittig, I., & Schägger, H. (2010). Large pore gels to separate mega protein complexes larger than 10MDa by blue native electrophoresis: Isolation of putative respiratory strings or patches. *Proteomics*, *10*(18), 3379–3387. https://doi.org/10.1002/pmic.201000343

Uliana, F., Ciuffa, R., Mishra, R., Fossati, A., Frommelt, F., Keller, S., Mehnert, M., Birkeland, E. S., van Drogen, F., Srejic, N., Peter, M., Tapon, N., Aebersold, R., & Gstaiger, M. (2023). Phosphorylation-linked complex profiling identifies assemblies required for Hippo signal integration. *Molecular Systems Biology*, *19*(4). https://doi.org/10.15252/msb.202211024

Van Strien, J., Guerrero-Castillo, S., Chatzispyrou, I. A., Houtkooper, R. H., Brandt, U., & Huynen, M. A. (2019). COmplexome Profiling ALignment (COPAL) reveals remodeling of mitochondrial protein complexes in Barth syndrome. *Bioinformatics*, 35(17), 3083–3091. https://doi.org/10.1093/bioinformatics/btz025

Wang, Y., Hu, Y., Höti, N., Huang, L., & Zhang, H. (2022). Characterization of In Vivo Protein Complexes via Chemical Cross-Linking and Mass Spectrometry. *Analytical Chemistry*, *94*(3), 1537–1542. https://doi.org/10.1021/acs.analchem.1c02410

Wittig, I., Braun, H. P., & Schägger, H. (2006). Blue native PAGE. *Nature Protocols*, *1*(1), 418–428. https://doi.org/10.1038/nprot.2006.62