

A Practical Beginner's Guide to Proteomics

This manuscript ([permalink](#)) was automatically generated from jessegmeyerlab/proteomics-tutorial@7f04f5d on September 22, 2023.

Authors

- **Oliver M. Crook**
 [0000-0001-5669-8506](#) ·  [ococrook](#) ·  [OllyMCrook](#)
Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom
- **Norbert Volkmar**
 [0000-0003-0766-5606](#) ·  [norbertvolkmar](#) ·  [NorbertVolkmar](#)
Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland; DISCO Pharmaceuticals, Zurich 8093, Switzerland
- **Susan B. Egbert**
 [0000-0001-5458-1099](#) ·  [lichenlady94](#) ·  [lichenlady94](#)
Department of Chemistry, University of Manitoba, Winnipeg, Canada
- **Benjamin A. Neely**
 [0000-0001-6120-7695](#) ·  [neely](#) ·  [neely615](#)
Chemical Sciences Division, National Institute of Standards and Technology, NIST Charleston · Funded by NIST
- **Germán L. Rosano**
 [0000-0002-8313-6813](#) ·  [ger225](#) ·  [GermanRosano](#)
Mass Spectrometry Unit, Institute of Molecular and Cellular Biology of Rosario, Rosario, Argentina · Funded by Grant PICT 2019-02971 (Agencia I+D+i)
- **Nicholas M. Riley**
 [0000-0002-1536-2966](#) ·  [rileynm](#) ·  [riley_nm1](#)
Department of Chemistry, University of Washington · Funded by Grant R00 GM147304
- **Martín L. Mayta**
 [0000-0002-7986-4551](#) ·  [martinmayta](#) ·  [MartinMayta2](#)
School of Medicine and Health Sciences, Center for Health Sciences Research, Universidad Adventista del Plata, Libertador San Martín 3103, Argentina; Molecular Biology Department, School of Pharmacy and Biochemistry, Universidad Nacional de Rosario, Rosario 2000, Argentina
- **Amit Kumar Yadav**
 [0000-0002-9445-8156](#) ·  [aky](#) ·  [theoneamit](#)
Translational Health Science and Technology Institute · Funded by Grant BT/PR16456/BID/7/624/2016 (Department of Biotechnology, India); Grant Translational Research Program (TRP) at THSTI funded by DBT
- **Devasahayam Arokia Balaya Rex**
 [0000-0002-9556-3150](#) ·  [ArokiaRex](#) ·  [rexprem](#)
Center for Systems Biology and Molecular Medicine, Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore 575018, India
- **Dina Schuster**
 [0000-0001-6611-8237](#) ·  [dschust-r](#) ·  [dina_sch](#)
Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich 8093, Switzerland; Department of

Biology, Institute of Molecular Biology and Biophysics, ETH Zurich, Zurich 8093, Switzerland; Laboratory of Biomolecular Research, Division of Biology and Chemistry, Paul Scherrer Institute, Villigen 5232, Switzerland

- **Jesse G. Meyer**

 [0000-0003-2753-3926](#) ·  [jessegmeyerlab](#) ·  [j_my_sci](#)

Department of Computational Biomedicine, Cedars Sinai Medical Center · Funded by Grant R21 AG074234; Grant R35 GM142502

- **Anna G. Duboff**

 [0009-0002-7316-3831](#) ·  [agduboff](#)

Department of Chemistry, University of Washington · Funded by Summer Research Acceleration Fellowship, Department of Chemistry, University of Washington

- **Emma H. Doud**

 [0000-0003-0049-0073](#) ·  [edoud1](#) ·  [fireinlab](#)

Center for Proteome Analysis, Indiana University School of Medicine, Indianapolis, Indiana, USA

- **Muralidharan Vanuopadath**

 [0000-0002-9364-917X](#) ·  [vanuopadathmurali](#) ·  [V_MuraleeDhar](#)

School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam-690 525, Kerala, India

✉ — Correspondence possible via [GitHub Issues](#)

Abstract

Proteomics is the large scale study of protein structure and function from biological systems. "Shotgun proteomics" or "bottom-up proteomics" is the prevailing strategy, in which proteins are hydrolyzed into peptide that are analyzed by mass spectrometry. Proteomics studies can be applied to diverse studies ranging from simple protein identification to studies of protein-protein interactions, absolute and relative protein quantification, post-translational modifications, and protein stability. To enable this range of different experiments, there are diverse strategies for proteome analysis. The nuances of how proteomic workflows differ may be difficult to understand for new practitioners. Here, we provide a comprehensive tutorial of different proteomics methods. Our tutorial covers all necessary steps starting from protein extraction and ending with biological interpretation. We expect that this work will serve as a basic resource for new practitioners of the field of shotgun or bottom-up proteomics.

Introduction

Proteomics is the large scale study of protein structure and function. Proteins are translated from mRNAs that are transcribed from the genome. Although the genome encodes potential cellular functions and states, the study of proteins is necessary to truly understand biology. Currently, proteomic studies are facilitated by mass spectrometry, although alternative methods are being developed.

Modern proteomics started around the year 1990 with the introduction of soft ionization methods that enabled, for the first time, transfer of large biomolecules into the gas phase without destroying them [1,2]. Shortly afterward, the first computer algorithm for matching peptides to a database was introduced [[PMID:24226387?](#)]. Another major milestone that allowed identification of over 1000 proteins were actually improvements to chromatography [3]. As the volume of data exploded, methods for statistical analysis transitioned use from the wild west to modern informatics based on statistical models [4] and the false discovery rate [5].

Two strategies of mass spectrometry-based proteomics differ fundamentally by whether proteins are cleaved into peptides before analysis: “top-down” and “bottom-up”. Bottom-up proteomics (also referred to as shotgun proteomics) is defined by the hydrolysis of proteins into peptide pieces [6]. Therefore, bottom-up proteomics does not actually measure proteins, but must infer their presence [4]. Sometimes proteins are inferred from only one peptide sequence representing a small fraction of the total protein sequence predicted from the genome. In contrast, top-down proteomics attempts to measure all proteins intact [7]. The potential benefit of top-down proteomics is the ability to measure proteoforms [8]. However, due to analytical challenges, the depth of protein coverage that is achievable by top-down proteomics is less than the depth that is achievable by bottom-up proteomics.

In this tutorial we focus on the bottom-up proteomics workflow. The most common version of this workflow is generally comprised of the following steps. First, proteins in a biological sample must be extracted. Usually this is done by denaturing and solubilizing the proteins while disrupting DNA and tissue. Next, proteins are hydrolyzed into peptides, usually using a protease like trypsin. Peptides from proteome hydrolysis must be purified. Most often this is done with reversed phase chromatography cartridges or tips. The peptides are then almost always separated by liquid chromatography before they are ionized and introduced into a mass spectrometer. The mass spectrometer then collects precursor and fragment ion data from those peptides. The data analysis is usually the rate limiting step. Peptides must be identified, and proteins are inferred and quantities are assigned. Changes in proteins across conditions are determined with statistical tests, and results must be interpreted in the context of the relevant biology.

There are many variations on this workflow. The wide variety of experimental goals that are achievable with proteomics technology leads to a wide variety of potential proteomics workflows. Even choice is important and every choice will affect the results. In this tutorial, we cover all of the required steps in detail to serve as a tutorial for new proteomics practitioners.

1. Biochemistry basics
2. Types of experiments enabled by proteomics
3. Protein extraction
4. proteolysis
5. Isotopic Labeling
6. Enrichments
7. Peptide purification
8. Mass Spectrometry

- 9. Peptide Ionization
- 10. Data Acquisition
- 11. Basic Data Analysis
- 12. Biological Interpretation
- 13. Sample Fractionation
- 14. Tandem MS
- 15. Experimental considerations and design

Biochemistry Basics

Proteins

Proteins are large biomolecules or biopolymers made up of amino acids which are linked by peptide bonds. They perform various functions in living organisms ranging from having structural roles to functional involvement in cellular signaling and the catalysis of chemical reactions (enzymes). Proteins are made up of 20 different amino acids (not counting pyrrolysine and selenocysteine, which only occur in specific organisms) and their sequence is encoded in their corresponding genes. The human genome encodes more than 20,000 different proteins. Each protein is present at a different abundances. Previous studies have shown that the concentration range of proteins can span over a range of at least seven orders of magnitude to up to 20 000 000 copies per cell and that their distribution is tissue-specific[[9]][10]. Due to genetic variations, as well as alternative splicing and post-translational modifications, multiple different proteoforms can be produced from one single gene[11].

PTMs

After protein biosynthesis, enzymatic and nonenzymatic processes change the protein sequence through proteolysis or chemical modification of the amino acid side chains. PTMs are an important biological process contributing to the diversity and control of the cellular proteome. Proteins can be post-translationally modified through enzymatic and non-enzymatic reactions *in vivo* and *in vitro* [12]. PTMs can be reversible or irreversible, and they change protein function in several ways, for example, by altering substrate–enzyme(s) interactions or altering protein-protein interactions [PMID:33826699?, PMID:24217768?].

More than 400 PTMs have been discovered in both prokaryotic and eukaryotic cells. These modifications are crucial in controlling protein functions and signal transduction pathways [PMID:23887885?]. The most commonly studied and biologically relevant post-translational modifications include ubiquitinations (Lys, Cys, Ser, Thr, N-term), succinylation (Lys), methylations (Arg, Lys, His, Glu, Asn, Cys), disulfide bonds (Cys-Cys), methylation (Arg, Lys), oxidations (any amino acid, but especially Met, Trp, His, Cys), phosphorylations (Ser, Thr, Tyr, His), acetylations (Lys, N-term), glycosylations (Arg, Asp, Cys, Ser, Thr, Tyr, Trp) and lipidations.

Post-translational modification of a protein can alter its function, activity, structure, location and interactions. PTMs alter signal transduction pathways and gene expression control [PMID:28656226?] regulation of apoptosis [PMID:23088365?, PMID:11368354?] by phosphorylation. Ubiquitination regulates protein degradation [PMID:16738015?], SUMOylation regulates chromatin structure, DNA repair, transcription, cell-cycle progression [PMID:26601932?, PMID:29079793?], and palmitoylation regulates maintenance of the structural organization of exosome-like extracellular vesicle membranes by [PMID:30251702?]. Glycosylation is a ubiquitous modification that regulates a variety of T cell functions, such as cellular migration, T cell receptor signalling, cell survival, and apoptosis [PMID:18846099?]. Deregulation of PTMs is linked to cellular stress and diseases [13].

Several non-MS methods exist to study PTMs, including *in vitro* PTM reaction tests with radioactive isotope-labelled substrates, western blot with PTM-specific antibodies, and peptide and protein arrays [PMID:12323352?]. While effective, these approaches have many limitations, such as inefficiency and difficulty in producing pan-specific antibodies. MS-based proteomics approaches are currently the predominant tool for identifying and quantifying changes in PTMs.

Protein Structure

Almost all proteins (except for intrinsically disordered proteins[14]) fold into 3D structures either by themselves or assisted through chaperones[15]. There are four levels relevant to the folding of any protein:

- Primary structure: The protein's linear amino acid sequence, with amino acids connected through peptide bonds.
- Secondary structure: The amino acid chain's folding: α -helix, β -sheet or turn.
- Tertiary structure: The three-dimensional structure of the protein.
- Quarternary structure: The structure of several protein molecules/subunits in one complex.

Biochemical and biophysical techniques for studying complex protein mixtures

Concentration/total protein amount determination

- UV-Vis Spectrophotometry (Bradford assay, BCA assay, Lowry, Folin)

Electrophoresis

- SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis)
- 2D-PAGE
- Isoelectric focussing (IEF)
- Western blotting

Biochemical and biophysical analytical techniques for studying purified proteins

Concentration determination

- UV-Vis Spectrophotometry

Protein size and aggregation

- Dynamic light scattering (DLS)
- Size-exclusion chromatography (SEC) or size-exclusion with multi angle laser light scattering (SEC-MALLS)
- SDS-PAGE
- Mass photometry

Structural techniques

- Circular dichroism (CD) spectropolarimetry
- X-ray crystallography
- Protein nuclear magnetic resonance spectroscopy (NMR)
- (cryo-)Electron microscopy (EM)

Interactions and binding kinetics

- Surface plasmon resonance (SPR)
- Isothermal titration calorimetry (ITC)

Buffers

Types of Experiments

A wide range of questions are addressable with proteomics technology, which translates to a wide range of variations of proteomics workflows. Sometimes identifying what proteins are present is desired, and sometimes the quantities of as many proteins as possible are desired. Proteomics experiments can be both qualitative and quantitative. The following sections give an overview of some common proteomics experiments.

Protein abundance changes

A common experiment is unbiased mapping of proteins along with detection of changes in their abundance across sample groups. This is achieved using one of the quantitative methods mentioned elsewhere in this manuscript such as LFQ or TMT. In these experiments, data should be collected from at least three biological replicates of each condition to estimate the variance of measuring each protein. Depending on the experiment design, different statistical tests are used to ask if there are changes in the measured protein abundances between groups. If there are only two groups, the quantities might be compared with a t-test or with a Wilcoxon signed-rank test. The latter is a non-parametric version of the t-test. If there are more than two sample groups, then often Analysis of Variance (ANOVA) is used instead. With either testing scheme, the p-values from the first set of tests must be corrected for multiple tests. A common method for p value correction is the Benjamini-Hochberg method [16]. These types of experiments have revealed wide ranges of proteomic remodeling from various biological systems.

Post-Translational Modifications (PTMs)

Proteins are modified after they are translated with various chemical groups, or even by proteolytic cleavage such as n-terminal methionine removal. There are several proteomics methods that are developed specific to each type of modification. See also the section on Protein/Peptide Enrichment and Depletion.

Phosphoproteomics

Phosphoproteomics is the study of the phosphorylation of proteins. Currently, there are two techniques that are useful for phosphoproteomics: mass spectrometry and western blotting. A key challenge with phosphoproteomics is sensitivity; it is important to ensure that there is sufficient amount of protein before completing a phosphoproteomics project, typically at least 1 mg of total protein per sample. After proteolysis of the proteome, phosphopeptides need to be enriched to be detected by mass spectrometry. Various methods of enrichment have been developed [17,18,19,20]. See the Peptide/Protein Enrichment and Depletion section for more details. Despite advancement in technology of phosphoproteomics, the following challenges still exist: limited sample amounts, highly complex samples, and huge dynamic range [21]. Additionally, phosphoproteomic analysis is often time-consuming and requires the use of expensive equipment such as enrichment kits.

Glycoproteomics

One PTM gaining interest due to its ubiquity and emerging functional roles is glycosylation. Protein glycosylation sites can be N-linked glycosylation sites (asparagine-linked), O-linked glycosylation sites (serine/threonine-linked). Understanding the function of protein glycosylation will help us understand numerous biological processes since this is a universal protein modification across all domains of life [22,23,24,25].

Studies of phosphorylation and glycosylation share several experimental pipeline steps including sample preparation. Protein clean-up approaches for glycoproteomics may differ from other proteomics experiments because glycopeptides are more hydrophilic than most peptides. Some approaches mentioned in literature include: filter-aided sample preparation (FASP), suspension traps (S-traps), and protein aggregation capture (PAC) [22,26,27,28,29,30,31]. Multiple proteases may be used to increase the sequence coverage and detect more modification sites, such as: trypsin, chymotrypsin, Pepsin, WalP/MaLP [32], GluC, AspN, Pronase, Proteinase K, OgpA, StcEz, BT4244, AM0627, AM1514, AM0608, Pic, ZmpC, CpaA, IMPa, PNGase F, Endo F, Endo H, and OglyZOR [22]. Mass spectrometry has improved over the past decade so that now more strategies are available for glycoprotein structure elucidation and also for glycosylation site quantification [22].

Structural techniques

Several proteomics methods have been developed to reveal protein structure information for simple and complex systems.

XL-MS (Cross-linking mass spectrometry)

XL-MS is an emerging technology for the field of proteomics. It can be used to determine changes in protein-protein interactions. XL-MS covalently locks interacting proteins together to preserve interactions and proximity during MS analysis. XL-MS is different from traditional MS in that it requires the identification of chimeric MS/MS spectra from cross-linked peptides [33,34].

The common steps in a XL-MS workflow are as follows [35]:

1. Generate a system with protein-protein interactions of interest (*in vitro* or *in vivo* [36])
2. Add a cross-linking reagent to covalently connect adjacent protein regions (such as disuccinimidyl sulfoxide, DSSO) [34]
3. Proteolysis to produce peptides
4. MS/MS data collection
5. Identify cross-linked peptide pairs using special software (i.e. pLink [37])
6. Generate cross-link maps for structural modeling

Hydrogen deuterium exchange mass spectrometry (HDX-MS)

HDX-MS works by detecting changes in peptide mass due to exchange of amide hydrogens of the protein backbone with deuterium from D₂O [38]. The rate of exchange depends on the protein solvent accessible surface area, dynamics, and the properties of the amino acid sequence [38,39,40, PMID:5333290?]. Although using D₂O as a solvent to make deuterium labeled sample is simple, HDX-MS requires a number of controls to ensure that experimental conditions capture the dynamics of interest [38,41,42,43]. If the peptide dissociation process is tuned appropriately, residue-level quantification of changes in solvent accessibility are possible within a measured peptide [44]. If done correctly, HDX can produce precise protein structure measurements with high reproducibility. Masson *et al.* gave recommendations on how to prep samples, conduct data analysis, and present findings in a detailed stepwise manner that will be useful to have if you are new to using this technique [38].

Radical Footprinting

This technique uses hydroxyl radical footprinting and MS in order to elucidate protein structures, assembly, and interactions within a large macromolecule [45,46]. In addition to proteomics applications, various approaches to make hydroxide radicals have also been applied for footprinting

studies in nucleic acid/ligand interactions [47,48,49]. This chapter is very useful in learning more on this topic [50].

There are several methods of producing radicals for protein footprinting:

1. Fenton and Fenton-like Chemistry [45,51,52]
2. Electron-Pulse Radiolysis [45,53]
3. High-Voltage Electrical Discharge [45,54]
4. Synchrotron X-ray Radiolysis of Water [45,55]
5. Plasma Formation of OH Radicals [45,56]
6. Photolysis of Hydrogen Peroxide [45,57]

Fast photochemical oxidation of proteins (FPOP) [58]

FPOP is an example of a radical footprinting method. In FPOP, a laser-based hydroxyl radical protein footprinting MS method that relies on the irreversible labeling of solvent-exposed amino acid side chains by hydroxyl radicals in order to understand structure of proteins. A laser produces 248 nm light that causes hydrogen peroxide to break into a pair of hydroxyl radicals [57,59]. The flow rate of solution through the capillary and laser frequency are adjusted so that each protein molecule is irradiated only once. After they are irradiated, the sample is collected to a tube that has catalase and free methionine in the buffer so that the H₂O₂ and hydroxyl radical are quenched to prevent further unwarranted oxidation, which may then modify residues that become exposed due unfolding after the initial labeling. Control samples are made by running the sample through the flow system without any irradiation. Another experimental control is done by adding a radical scavenger to tune the extent of protein oxidation [60,61]. FPOP has wide application for proteins including measurements of fast protein folding and transient dynamics.

Protein Painting [62,63]

Protein painting uses “molecular paints” to noncovalently coat the solvent accessible surface of proteins. These paint molecules will coat the protein surfaces but will not have access to the hydrophobic cores or protein-protein interface regions that solvents cannot access. If the “paint” covers free amines of lysine side chains, the parts that are painted will not be subjected to trypsin cleavage, while the unpainted areas will. After proteolysis, the peptides samples will be subjected to MS. A lack of proteolysis in a region is interpreted as solvent accessibility, which gives rough structural information about complex protein mixtures or even a whole proteome.

LiP-MS (limited proteolysis mass spectrometry) [64,65,65,66]

Limited proteolysis coupled to mass spectrometry (LiP-MS) is a method that tracks structural changes in complex proteomes in response to arbitrary stimuli. The idea is to detect changes in protease-accessible protein regions due to the perturbation. For example, a drug binding may make a disordered region protected from non-specific proteolysis. LiP-MS can therefore provide a somewhat unbiased view of structural changes at the proteome scale. This may be useful when attempting to discover a mechanism for some treatment that causes a phenotype. For additional information about LiP-MS, please refer to this document: [67]

Protein stability and small molecule binding

Thermal proteome profiling (TPP) [68,69,70,71]

Thermal proteome profiling (TPP) involves subjecting a protein sample to a thermal shift assay (TSA), in which the protein is exposed to a range of temperatures, and the resulting changes in protein

stability are monitored using mass spectrometry. During a typical TPP experiment, a protein sample is first treated with a drug of interest to stabilize protein-ligand interactions. The sample is then divided into multiple aliquots, which are subjected to different temperatures in order to induce thermal denaturation. The resulting drug induced changes in protein stability curves are detected using mass spectrometry. By comparing protein stability curves across the temperatures between treatment conditions, TPP can provide insight into the proteins that bind a ligand.

Cellular Thermal Shift Assay (CETSA) [72]

CETSA is the same idea as TPP, but specifically for studies of protein stability in live cells. The assay is capable of detecting shifts in the thermal equilibrium of cellular proteins in response to a variety of perturbations, but most commonly in response to drug treatments.

Protein-protein interactions (PPIs)

Affinity purification coupled to mass spectrometry (AP-MS) [73,74,75]

AP-MS is an approach that involves the purification of a target protein or protein complex using a specific antibody followed by mass spectrometry analysis to identify the interacting proteins. In a typical AP-MS experiment, a protein or protein complex of interest is first tagged with a specific epitope or affinity tag, such as a FLAG or HA tag, which is used to selectively capture the target protein using an antibody. The protein complex is then purified from the sample using a series of wash steps, and the interacting proteins are identified using mass spectrometry. The success of AP-MS experiments depends on a number of factors, including the quality of the antibody or tag used for purification, the specificity and efficiency of the resin used for capture, and the sensitivity and resolution of the mass spectrometer. In addition, careful experimental design and data analysis are critical for accurate identification and interpretation of protein-protein interactions. AP-MS has been used to study a wide range of biological processes, including signal transduction pathways, protein complexes, and protein post-translational modifications. AP-MS has been performed on a whole proteome scale as part of the BioPlex project [76,77,78]. Despite its widespread use, AP-MS has some limitations, including the potential for non-specific interactions, the difficulty in interpreting complex data sets, and the possibility of missing important interacting partners due to limitations in sensitivity or specificity. However, with continued advances in technology and data analysis methods, AP-MS is likely to remain a valuable tool for the study of protein-protein interactions.

APEX peroxidase [79,80]

APEX-MS is a labeling technique that enables genetically attaches a peroxidase to a protein of interest. When biotin-phenol is transiently added in the presence of hydrogen peroxide, nearby proteins are covalently biotinylated [81]. APEX thereby enables discovery of interacting proteins in living cells. One of the major advantages of APEX is its ability to label proteins in their native environment, allowing for the identification of interactions that occur under physiological conditions. Despite its advantages, APEX has some limitations, including the potential for non-specific labeling, the difficulty in distinguishing between direct and indirect interactions, and the possibility of missing interactions that occur at low abundance or in regions of the cell that are not effectively labeled.

Proximity-dependent biotin identification (BioID) [82,83,84,85]

BioID is a proximity labeling technique that allows for the identification of protein-protein interactions. BioID involves the genetic tagging of a protein of interest with a promiscuous biotin ligase in live cells, which then biotinylates proteins in close proximity to the protein of interest. One of the advantages of BioID is its ability to label proteins in their native environment, allowing for the identification of

interactions that occur under physiological conditions. BiOID has been used to identify a wide range of protein interactions, including receptor-ligand interactions, signaling complexes, and protein localization. BiOID is a slower reaction than APEX and therefore may pick up more transient interactions. BiOID has the same limitations as APEX. For more information on BiOID, please refer to this chapter: [86]

Protein Extraction

Protein extraction from the sample of interest is the initial phase of any mass spectrometry-based proteomics experiment. Thought should be given to any planned downstream assays, specific needs of proteolysis (LiP-MS, post translational modification enrichments, enzymatic reactions, glycan purification or hydrogen-deuterium exchange experiments) long term project goals (reproducibility, multiple sample types, low abundance samples) as well as to the initial experimental question (coverage of a specific protein, subcellular proteomics, global proteomics, protein-protein interactions or immune or affinity enrichment of a specific classes of modifications.) The 2009 version of Methods in Enzymology: guide to Protein Purification [87] serves as a deep dive into how molecular biologists and biochemists traditionally thought about protein extraction. The Protein Protocols handbook [88] and the excellent review by Linn [[PMID:19892162?](#)] are good sources of general proteomics protocols for a scientist new to the field. Any change in extraction conditions should be expected to create potential changes in downstream results. Be sure to think about and optimize the protein extraction step first and stick with a protocol that works for your needs. If a collaborator is attempting to reproduce your results, make sure they begin with the same extraction protocols.

Buffer choice

General proteomics

A common question to proteomics core facilities is, "What is the best buffer for protein extraction?" Unfortunately, there is no one correct answer. For global proteomics experiments where maximizing the number of protein or peptide identifications is a goal, a buffer of neutral pH (50-100 mM PBS, Tris, HEPES, ammonium bicarbonate, triethanolamine bicarbonate; pH 7.5-8.5) is used in conjunction with a chaotrope or surfactant to denature and solubilize proteins (e.g., 8 M urea, 6 M guanidine, 5% SDS) [[PMID:16152629?](#),[PMID:20722421?](#)]. Often other salts like 50-150 mM NaCl are also added. Complete denaturation of the proteins in the sample in a timely fashion is an advantage as it generally prevents changes to protein status by endogenous proteases, kinases, phosphatases, and other enzymes. If intact protein separations are planned (based on size or isoelectric point) choose a denaturant compatible with those methods, such as SDS [[PMID:31249407?](#)]. Compatibility with protease (typically trypsin) and peptide cleanup steps will need to be considered. 8 M urea must be diluted to 2 M or less for trypsin and chymotrypsin digestions, while guanidine and SDS should be removed either through protein precipitation, through filter-assisted sample preparation (FASP), or similar solid phase digestion techniques. Note that some buffers can potentially introduce modifications onto proteins such as carbamylation from urea at high temperatures [[PMID:24161613?](#)].

Protein-protein interactions

Denaturing conditions will efficiently extract proteins – but they will denature/disrupt most protein-protein interactions. If you are working on an immune- or affinity purification of a specific protein and expect to analyze enzymatic activity, structural features, and/or protein-protein interactions, a non-denaturing lysis buffer should be utilized [[PMID:21364760?](#),[PMID:10504710?](#)]. Check the calculated pI and hydrophobicity (the Expasy.org resource ProtParam is useful for this) for a good idea of starting pH/conductivity, but you may need to perform a stability screen. In general, a good starting point for the buffer will still be close to neutral pH with 50-250 mM NaCl, but specific proteins may require pH as low as 2 or as high as 9 for stable extraction. A low percent of mass spec compatible detergent may also be used. Newer mass spectrometry compatible detergents are also useful for protein extraction and ease of downstream processing – including Rapigest® (Waters), N-octyl-β-glucopyranoside, Azo [[PMID:33232116?](#)], PPS silent surfactant [[PMID:21280217?](#)], sodium laurate [[PMID:23555778?](#)], and

sodium deoxycholate [[PMID:17022626?](#)]. AVOID the use of tween-20, triton-X, NP-40, and PEGs as these compounds are challenging to remove after digestion [[PMID:29726681?](#)].

Optional additives

For non-denaturing buffer conditions, which preserve tertiary and quaternary protein structures, additional additives may not be necessary for successful extraction and to prevent proteolysis or PTM modifications throughout the extraction process. Protease, phosphatase and deubiquitinase inhibitors are optional additives in less denaturing conditions or in experiments focused on specific post-translational modifications. Keep in mind that protease inhibitors may impact digestion conditions and will need to be diluted or removed prior to trypsin addition. For extraction of DNA or RNA binding proteins, addition of a small amount of nuclease or benzonase might be useful for degradation of any bound nucleic acids and result in a more consistent digestion [[PMID:23792921?](#)].

Mechanical or Sonic Disruption

Cell lysis

One typical lysis buffer is 8 M urea in 100 mM Tris, pH 8.5; the pH based on optimum trypsin activity [[PMID:25664860?](#)] Small mammalian cell pellets and exosomes will lyse almost instantly upon addition denaturing buffer. If non-denaturing conditions are desired, osmotic swelling and subsequent shearing or sonication can be applied [89]. Efficiency of extraction and degradation of nucleic acids can be improved using various sonication methods: 1) probe sonicator with ice; 2) water bath sonicator with ice or cooling; 3) bioruptor® sonication device 4) Adaptive focused acoustics (AFA®) [[PMID:21060726?](#)]. Key to these additional lysis techniques are to keep the temperature of the sample from rising significantly which can cause proteins to aggregate or degrade. Some cell types may require additional force for effective lysis (see below). For cells with cell walls (i.e. bacteria or yeast), lysozyme is often added in the lysis buffer. Any added protein will be present in downstream results, however, so excessive addition of lysozyme is to be avoided unless tagged protein purification will occur.

Tissue/other lysis

Although small pieces of soft tissue can often be successfully extracted with the probe and sonication methods described above, larger/harder tissues as well as plants/yeast/fungi are better extracted with some form of additional mechanical force. If proteins are to be extracted from a large amount of sample, such as soil, feces, or other diffuse input, one option is to use a dedicated blender and filter the sample, followed by centrifugation. If samples are smaller, such as tissue, tumors, etc., cryo-homogenization is recommended. The simplest form of this is grinding the sample with liquid nitrogen and a mortar and pestle. Tools such as bead beaters (i.e. FastPrep-24®) are also used, where the sample is placed in a tube with appropriately sized glass or ceramics beads and shaken rapidly. Cryo-mills are chambers where liquid nitrogen is applied around a vessel and large bead or beads. Cryo-fractionators homogenize samples in special bags that are frozen in liquid nitrogen and smashed with various degrees of force [[PMID:34002278?](#)]. After homogenization, samples can be sonicated by one of the methods above to fragment DNA and increase solubilization of proteins.

Measuring the efficiency of protein extraction

Following protein extraction, samples should be centrifuged (10-14,000 g for 10-30 min depending on sample type) to remove debris and any unlysed material prior to determining protein concentration. The amount of remaining insoluble material should be noted throughout an experiment as a large change may indicate protein extraction issues. Protein concentration can be calculated using a

number of assays or tools [[PMID:18429326?](#), [PMID:12703310?](#)]; generally absorbance measurements are facile, fast and affordable, such as Bradford or BCA assays. Protein can also be estimated by tryptophan fluorescence, which has the benefit of not consuming sample [90]. A nanodrop UV spectrophotometer may be used to measure absorbance at UV280. Consistency in this method is important as each method will have inherent bias and error [[PMID:26342307?](#), [PMID:30234128?](#)]. Extraction buffer components will need to be compatible with any assay chosen; alternatively, buffer may be removed (see below) prior to protein concentration calculation.

Reduction and alkylation

Typically, disulfide bonds in proteins are reduced and alkylated prior to proteolysis in order to disrupt structures and simplify peptide analysis. This allows better access to all residues during proteolysis and removes the crosslinked peptides created by S-S inter peptide linkages. There are a variety of reagent options for these steps. For reduction, the typical agents used are 5-15 mM concentration of tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl), dithiothreitol (DTT), or 2-mercaptoethanol (2BME). TCEP-HCl is an efficient reducing agent, but it also significantly lowers sample pH, which can be abated by increasing sample buffer concentration or resuspending TCEP-HCl in an appropriate buffer system (i.e 1M HEPES pH 7.5).

Following the reducing step, a slightly higher 10-20mM concentration of alkylating agent such as chloroacetamide/iodoacetamide or n-ethyl maleimide is used to cap the free thiols [[PMID:29019370?](#); [\[PMID:15351294?\]](#); [\[PMID:28539326?\]](#)]. In order to monitor which cysteine residues are linked or modified in a protein, it is also possible to alkylate free cysteines with one reagent, reduce di-sulfide bonds (or other cysteine modifications) and alkylate with a different reagent [[PMID:32132231?](#), [PMID:28445428?](#), [PMID:23074338?](#)]. Alkylation reactions are generally carried out in the dark at room temperature to avoid excessive off-target alkylation of other amino acids.

Removal of buffer/interfering small molecules

If extraction must take place in a buffer which is incompatible for efficient proteolysis (check the guidelines for the protease of choice), then protein cleanup should occur prior to digestion. This is generally performed through precipitation of proteins. The most common types are 1) acetone, 2) trichloroacetic acid (TCA), and 3) methanol/chloroform/water [[PMID:14753699?](#), [PMID:19892180?](#)]. Proteins are generally insoluble in most pure organic solvents, so cold ethanol or methanol are sometimes used. Pellets should be washed with organic solvent for complete removal especially of detergents. Alternatively, solid phase based digestion methods such as S-trap [[PMID:33750040?](#)], FASP [[PMID:19377485?](#), [PMID:30259475?](#)], SP3 [[PMID:3117935?](#), [PMID:28948796?](#)] and on column/bead can allow for proteins to be applied to a solid phase and buffers removed prior to proteolysis [[PMID:29754492?](#)]. Specialty detergent removal columns exist (Pierce/Thermo Fisher Scientific) but add expense and time consuming steps to the process. Relatively low concentrations of specific detergents, such as 1% deoxycholate (DOC), or chaotropes (i.e. 1M urea) are compatible with proteolysis by trypsin/Lys-C. Often proteolysis-compatible concentrations of these detergents and chaotropes are achieved by diluting the sample in appropriate buffer (i.e. 100 mM ammonium bicarbonate, pH 8.5) after cell or tissue lysis in a higher concentration. DOC can then be easily removed by precipitation or phase separation [[PMID:18183947?](#)] following digestion by acidification of the sample to pH 2-3.

Any small-molecule removal protocol should be tested for efficiency prior to implementing in a workflow with many samples as avoiding detergent (or polymer) contamination in the LC/MS is very important.

Protein quantification

After proteins are isolated from the sample matrix, they are often quantified. Protein quantification is important to assess the yield of an extraction procedure, and to adjust the scale of the downstream processing steps to match the amount of protein. For example, when purifying peptides, the amount of sorbent should match the amount of material to be bound. Presently, there is a wide variety of techniques to quantitate the amount of protein present in a given sample. These methods can be broadly divided into three types as follows:

Colorimetry-based methods:

The method includes different assays like Coomassie Blue G-250 dye binding (the Bradford assay), the Folin-Lowry assay, the bicinchoninic acid (BCA) assay and the biuret assay [[PMID:10075906?](#)]. The most commonly used method is the BCA assay. In the BCA method the peptide bonds of the protein reduce cupric ions $[Cu^{2+}]$ to cuprous ions $[Cu^{+}]$ at a rate which is proportional to the amount of protein present in the sample. Subsequently, the BCA reagent binds to the cuprous ions, leading to the formation of a complex which absorbs 562 nm wavelength light. This permits a direct correlation between sample protein concentration and absorbance [[PMID:3843705?](#), [PMID:7951748?](#)]. The Bradford assay is another method for protein quantification also based on colorimetry principle. It relies on the interaction between the Coomassie brilliant blue dye and the protein based on hydrophobic and electrostatic interactions. Dye binding shifts the absorption maxima from 470 nm to 595 nm [[PMID:32238597?](#)]. Similarly, the Folin-Lowry method is a two-step colorimetric assay. Step one is the biuret reaction wherein complexes of copper with the nitrogen in the protein molecule are formed. In the second step, the complexed tyrosine and tryptophan amino acids react with Folin-Ciocalteu phenol reagent generating an intense, blue-green color absorbing light at 650–750 nm [[PMID:6744121?](#)].

Another simple but less reliable protein quantification method of UV-Vis Absorbance at 280 nm estimates the protein concentration by measuring the absorption of the aromatic residues; tyrosine, and tryptophan, at 280 nm [[PMID:34533299?](#)].

Fluorescence-based methods:

Colorimetric assays are inexpensive and require common lab equipment, but colorimetric detection is less sensitive than fluorescence. Protein in proteomic samples can be quantified using intrinsic fluorescence of tryptophan based on the assumption that approximately 1% of all amino acids in the proteome are tryptophan [[PMID:25837572?](#)].

NanoOrange is an assay for the quantitative measurement of proteins in solution using the NanoOrange reagent, a merocyanine dye that produces a large increase in fluorescence quantum yield when it interacts with detergent-coated proteins. Fluorescence is measured using 485-nm excitation and 590-nm emission wavelengths. The NanoOrange assay can be performed using fluorescence microplate readers, fluorometers, and laser scanners that are standard in the laboratory [[PMID:12703310?](#)].

3-(4-carboxybenzoyl)quinoline-2-carboxaldehyde (CBQCA) is a sensitive fluorogenic reagent for amine detection, which can be used for analyzing proteins in solution. As the number of accessible amines in a protein is modulated by its concentration, CBQCA has a greater sensitivity and dynamic range when measuring protein concentration [[PMID:9025944?](#)].

Summary

Often you will be given protein extraction conditions from molecular biologists or biochemistry which you will have to make work with downstream mass spectrometry applications. For bottom-up

proteomics, the overreaching goal is efficient and consistent extraction and digestion. A range of mechanical and non-mechanical extraction protocols have been developed and the use any one specific technique is generally dictated by sample type or assay requirements (i.e. native versus non-native extraction). Extraction can be aided by the addition of detergents and/or chaotropes to the sample, but care should be taken that these additives do not interfere with the sample digestion step or downstream mass-spectrometry applications.

Proteolysis

Proteolysis is the defining step that differentiates bottom-up or shotgun proteomics from top-down proteomics. Hydrolysis of proteins is extremely important because it defines the population of potentially identifiable peptides. Generally peptides between a length of 7-35 amino acids are considered useful for mass spectrometry analysis. Peptides that are too long are difficult to identify by tandem mass spectrometry, or may be lost during sample preparation due to irreversible binding with solid-phase extraction sorbents. Peptides that are too short are also not useful because they may match to many proteins during protein inference. There are many choices of enzymes and chemicals that hydrolyze proteins into peptides. This section summarizes potential choices and their strengths and weaknesses.

Trypsin is the most common choice of protease for proteome hydrolysis [91]. Trypsin is favorable because of its specificity, availability, efficiency and low cost. Trypsin cleaves at the C-terminus of basic amino acids, Arg and Lys. Many of the peptides generated from trypsin are short in length (less than ~20 amino acids), which is ideal for chromatographic separation, MS-based peptide fragmentation and identification by database search. The main drawback of trypsin is that majority (56%) of the tryptic peptides are ≤ 6 amino acids, and hence using trypsin alone limits the observable proteome [PMID:20113005?, PMID:25823410?, PMID:30687733?]. This limits the number of identifiable protein isoforms and post-translational modifications.

3. theoretical studies of proteolysis enzymes [92]

4. Challenges associated with alternative enzyme choices (non-specific and semi-specific enzymes)

Many alternative proteases are available with different specificities that complement trypsin to reveal different proteomic sequences [PMID:12643544?, PMID:20113005?], which can help distinguish protein isoforms [PMID:27123950?]. The enzyme choice mostly depends on the application. In general, for a mere protein identification mostly trypsin is the choice due to the reasons aforementioned. However, alternative enzymes can facilitate *de novo* assembly when the genomic data information is limited in the public database repositories

[93,94,pmid:31615963?,pmid:30622160?,pmid:29990557?]. Use of multiple proteases for proteome digestion also can improve the sensitivity and accuracy of protein quantification [PMID:30336047?]. Moreover, by providing an increased peptide diversity, the use of multiple proteases can expand sequence coverage and increase the probability of finding peptides which are unique to single proteins [32,92,95]. A multi-protease approach can also improve the identification of N-Termini and signal peptides for small proteins [96]. Overall, integrating multiple-protease data can increase the number of proteins identified [97,98], the number of identified post-translational modifications detected [32,95,99] and decrease the ambiguity of the protein group list [95].

Lysyl endopeptidase (Lys-C) obtained from *Lysobacter enzymogenes* is a serine protease involved in cleaving carboxyl terminus of Lys [PMID:25823410?]. Like trypsin, the optimum pH range required for its activity is from 7 to 9. A major advantage of Lys-C is its resistance to denaturing agents, including 8 M urea - a chaotrope commonly used to denature proteins *prior* to digestion [PMID:27123950?]. Trypsin is less efficient at cleaving Lys than Arg, which could limit the quality of quantitation from tryptic peptides. Hence, to achieve complete protein digestion with minimal missed cleavages, Lys-C is often used simultaneously with trypsin digestion [PMID:23017020?].

Alpha-lytic protease (aLP) is also secreted by the soil bacterial *Lysobacter enzymogenes* [PMID:3053694?]. Wild-type aLP (WaLP) and an active site mutant of aLP, M190A (MaLP), have been used to expand proteome coverage [32]. Based on observed peptide sequences from yeast proteome digestion, WaLP showed a specificity for small aliphatic amino acids like alanine, valine, and glycine, but also threonine and serine. MaLP showed specificity for slightly larger amino acids like methionine,

phenylalanine, and surprisingly, a preference for leucine over isoleucine. The specificity of WaLP for threonine enabled the first method for mapping endogenous human SUMO sites [[PMID:29079793?](#)].

Glutamyl peptidase I, commonly known as Glu-C or V8 protease, is a serine protease obtained from *Staphylococcus aureus* [[PMID:4627743?](#)]. Glu-C cleaves at the C-terminus of glutamate, but also after aspartate [[PMID:4627743?](#), [PMID:26748652?](#)].

Peptidyl-Asp metallopeptidase, commonly known as Asp-N, is a metalloprotease obtained from *Pseudomonas fragi* [[PMID:2669754?](#)]. Asp-N catalyzes the hydrolysis of peptide bonds at the N-terminal of aspartate residues. The optimum activity of this enzyme occurs at a pH range between 4 and 9.

As with any metalloprotease, chelators like EDTA should be avoided for digestion buffers when using Asp-N. Studies also suggest that Asp-N cleaves at the amino terminus of glutamate when a detergent is present in the proteolysis buffer [[PMID:2669754?](#)]. Asp-N often leaves many missed cleavages [[PMID:27123950?](#)].

Chymotrypsin or chymotrypsinogen A is a serine protease obtained from porcine or bovine pancreas with an optimum pH range from 7.8 to 8.0 [[PMID:3555886?](#)]. It cleaves at the C-terminus of hydrophobic amino acids Phe, Trp, Tyr and barely Met and Leu residues. Since the transmembrane region of membrane proteins commonly lacks tryptic cleavage sites, this enzyme works well with membrane proteins having more hydrophobic residues [[PMID:24870543?](#), [PMID:24696503?](#), [PMID:27123950?](#)]. The chymotryptic peptides generated after proteolysis will cover the proteome space orthogonal to that of tryptic peptides both in a quantitative and qualitative manner [[PMID:24290761?](#), [PMID:22669647?](#), [PMID:24696503?](#)]

Clostripain, commonly known as Arg-C, is a cysteine protease obtained from *Clostridium histolyticum* [[PMID:4332560?](#)]. It hydrolyses mostly the C-terminal Arg residues and sometimes Lys residues, but with less efficiency. The peptides generated are generally longer than that of tryptic peptides. Arg-C is often used with other proteases for improving qualitative proteome data and also for investigating PTMs [[PMID:25823410?](#)].

LysargiNase, also known as Ulilysin, is a recently discovered protease belonging to the metalloprotease family. It is a thermophilic protease derived from *Methanosaerina acetivorans* that specifically cleaves at the N-terminus of Lys and Arg residues [[PMID:25419962?](#)]. Hence, it enabled discovery of C-terminal peptides that were not observed using trypsin. In addition, it can also cleave modified amino acids such as methylated or dimethylated Arg and Lys [[PMID:25419962?](#)].

Peptidyl-Lys metalloendopeptidase, or Lys-N, is an metalloprotease obtained from *Grifola frondosa* [[PMID:19195997?](#)]. It cleaves N-terminally of Lys and has an optimal activity at pH 9.0. Unlike trypsin, Lys-N is more resistant to denaturing agents and can be heated up to 70 °C [[PMID:25823410?](#)]. Reports suggest that the peptides generated after Lys-N digestion produces more of c-type ions in a ETD-based mass spectrometer [[PMID:18425140?](#)]. Hence this can be used for analysing PTMs, identification of C-terminal peptides and also for *de novo* sequencing strategies [[PMID:18425140?](#), [PMID:20953479?](#)].

Pepsin A, commonly known as pepsin, is an aspartic protease obtained from bovine or porcine pancreas [[PMID:12089768?](#)]. Pepsin was one of several proteins crystallized by John Northrop, who shared the 1946 Nobel prize in chemistry for this work [[100/](#), [PMID:19872561?](#), [PMID:19872562?](#), [PMID:17758437?](#)]. Pepsin works at an optimum pH range from 1 to 4 and specifically cleaves Trp, Phe, Tyr and Leu [[PMID:25823410?](#)]. Since it possess high enzyme activity and broad specificity at lower pH, it is preferred over other proteases for MS-based disulphide mapping [[PMID:12476442?](#), [PMID:24980484?](#)]. Pepsin is also used extensively for structural

mass spectrometry studies with hydrogen-deuterium exchange (HDX) because the rate of back exchange of the amide deuteron is minimized at low pH [101,102].

Proteinase K was first isolated from the mold *Tritirachium album* Limber [PMID:4373242?]. The epithet 'K' is derived from its ability to efficiently hydrolyse keratin [PMID:4373242?]. It is a member of the subtilisin family of proteases and is relatively unspecific with a preference for proteolysis at hydrophobic and aromatic amino acid residues [103]. The optimal enzyme activity is between pH 7.5 and 12. Proteinase K is used at low concentrations for limited proteolysis (LiP) and the detection of protein structural changes in the eponymous technique LiP-MS [PMID:29072706?].

Multiple-protease-based proteomic analysis

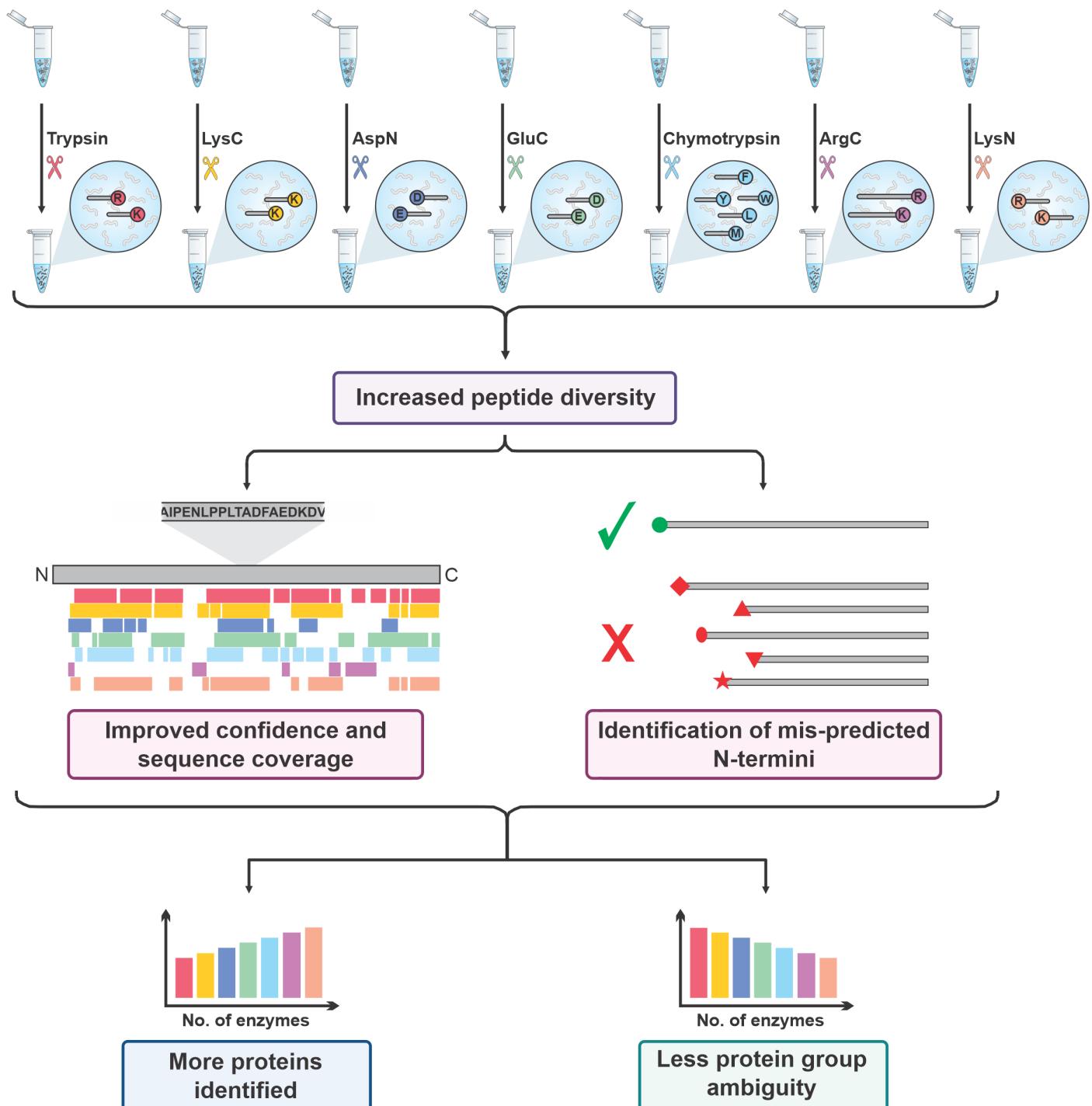


Figure 1: Multiple protease proteolysis improves protein inference The use of other proteases beyond Trypsin such as Lysyl endopeptidase (Lys-C), Peptidyl-Asp metallopeptidase (Asp-N), Glutamyl peptidase I, (Glu-C), Chymotrypsin, Clostripain (Arg-C) or Peptidyl-Lys metalloendopeptidase (Lys-N) can generate a greater diversity of peptides. This improves protein sequence coverage and allows for the correct identification of their N-termini. Increasing the number

of complimentary enzymes used will increase the number of proteins identified by single peptides and decreases the ambiguity of the assignment of protein groups. Therefore, this will allow more protein isoforms and post-translational modifications to be identified than using Trypsin alone.

Peptide Quantification

Discussion of methods to isotopically label peptides or proteins that enable quantification

1. SILAC/SILAM
2. dimethyl labeling
3. Isobaric tags

Peptide labeling with isobaric tags

The isobaric tag labeling-based quantitation uses derivatization of every peptide sample with a different isotope from a set of isobaric mass tags. This is followed by pooling the labelled samples, which undergo MS analysis simultaneously. As the isobaric tags are used, peptides labeled with these tags give a single MS peak with the same precursor m/z value in an MS1 scan and identical retention time of liquid chromatography analysis. The modified parent ions undergo fragmentation during MS/MS analysis generating two kinds of fragment ions: (a) reporter ions and (b) peptide fragment ions. The reporter ions' relative intensity is directly proportional to the amount of peptide in each of the starting samples that were mixed. The fragment ion peaks with higher m/z values correspond to amino acid sequences of peptides and are used for identifying peptides, from which proteins can be inferred. Since it is possible to label most tryptic peptide with an isobaric mass tag at least at the N-termini, numerous peptides from the same protein can be detected, thus leading to an increase in the confidence in both protein identification and quantification [[PMID:25337643?](#)]. All isobaric tags have a common structural theme consisting of 1) an amine-reactive group (usually triazine ester or N-hydroxysuccinimide [NHS] esters) which react with peptide N-termini and ε-amino group of the lysine side chain of peptides, 2) a balancer group, and 3) a reporter ion group.

Because the size of the reporter ions is small and sometimes the mass difference between reporter ions is small, these methods are mostly used with high-resolution mass measurement, not with classical ion traps [[PMID:26584918?](#)]. There are examples, however, of using isobaric tags with pulsed q dissociation on linear ion traps (LTQs) [[PMID:22397766?](#)]. Suitable instruments are the Thermo Q-Exactive, Exploris lines, and Tribrid lines, or TOFs such as the TripleTOF or timsTOFs [[104, PMID:30967486?](#)].

The following are some of the isobaric labeling techniques:

isobaric Tags for Relative and Absolute Quantitation (iTRAQ)

The iTRAQ tagging method covalently labels the peptide N-terminus and side-chain primary amines with tags of different masses through the NHS-ester bond. This is followed by mass spectrometry analysis [[PMID:15385600?](#)]. Reporter ions for an 8plex iTRAQ are measured at roughly 113, 114, 115, 116, 117, 118, 119, and 121 Thompsons. At the moment, two kinds of iTRAQ reagents are available: 4-plex and 8-plex. Using 4-plex reagents, a maximum of four different biological conditions can be analyzed simultaneously, whereas using 8plex reagents, eight different biological conditions can be analysed [[PMID:20593797, PMID:22594965?](#)].

iTRAQ hydrazide (iTRAQH)

iTRAQH is an isobaric tagging reagent for the selective labeling and relative quantification of carbonyl (CO) groups in proteins [[PMID:22926130?](#)]. The reactive CO and oxygen groups which are generated as the byproducts of oxidation of lipids at the time of oxidative stress causes protein carbonylation [[PMID:15775985?](#)]. iTRAQH is produced from iTRAQ and surplus of hydrazine. This reagent reacts with

peptides which are carbonylated, thus forming a hydrazone group. iTRAQH is a novel method for analyzing carbonylation sites in proteins utilizing an isobaric tag for absolute and relative quantitation iTRAQ derivative, iTRAQH, and the analytical power of linear ion trap instruments (QqLIT). This new strategy seems to be well suited for quantifying carbonylation at large scales because it avoids time-consuming enrichment procedures [[PMID:22926130?](#)]. Thus, there is no need for enriching modified peptides before LC-MS/MS analysis.

Tandem Mass Tag (TMT)

TMT labeling is based on a similar principle as that of iTRAQ. In the case of 6-plex-TMT, the masses of reporter groups are roughly 126, 127, 128, 129, 130, and 131 Thompsons [[PMID:26584918?](#)]. TMT works best with MS which allow quantitation at MS3-level with higher accuracy (e.g.: Thermo's Fusion Orbitrap instruments) [[PMID:25337643?](#)] and its eliminates ratio distortion in isobaric multiplexed quantitative proteomics [[PMID:21963607?](#)]. The TMT is widely used for quantitative protein biomarker discovery. In addition, TMT labeling technique helps multiplex sample analysis enabling efficient use of instrument time. TMT labelling also controls for technical variation because after samples are mixed the ratios are locked in, and any sample loss would be equal across channels. A wide range of TMT reagents with different multiplexing capabilities are available, such as TMT zero, TMT duplex, TMT six plex, TMT 10-plex, and TMT 11-plex are available along with the recent addition of TMT 16-plex and now even TMT 18-plex [[105](#)]. These TMT reagents have a similar chemical structure, which allows the efficient transition from method development to multiplex quantification [[PMID:30967486?](#)].

iodoTMT

IodoTMT reagents are isobaric reagents used for tagging cysteine residues of peptides. The commercially available IodoTMT reagents are iodoTMTzero and iodoTMT 6-plex [[PMID:24152285?](#), [PMID:24926564?](#)].

aminoxyTMT Isobaric Mass Tags

Also referred to as glyco-TMTs, these reagents have chemistry similar to iTRAQH. The stable isotope-labeled glyco-TMTs are utilized for quantitating N-linked glycans. They are derived from the original TMT reagents with an addition of carbonyl-reactive groups, which involve either hydrazide or aminoxy chemistry as functional groups. These aminoxy TMTs show a better performance as compared to its iTRAQH counterparts in terms of efficiency of labeling and quantification. The glyco-TMT compounds consist of stable isotopes thus enabling (i) isobaric quantification using MS/MS spectra and (ii) quantification in MS1 spectra using heavy/light pairs. Aminoxy TMT6-128 and TMT6-131 along with the hydrazide TMT2-126 and TMT2-127 reagents can be used for isobaric quantification. In the quantification at MS1 level, the light TMT0 and the heavy TMT6 reagents have a difference in mass of 5.0105 Da which is sufficient to separate the isotopic patterns of all common N-glycans. Glycan quantification based on glyco-TMTs generates more accurate quantification in MS1 spectra over a broad dynamic range. Intact proteins or their digests obtained from biological samples are treated with PNGase F/A glycosidases to release the N-linked glycans during the process of labeling using aminoxyTMT reagents. The free glycans are then purified and labeled with the aminoxyTMT reagent at the reducing end. The labeled glycans from individual samples are subsequently pooled and then undergo analysis in MS for identification of glycoforms in the sample and quantification of relative abundance of reporter ions at MS/MS level [[PMID:22455665?](#)].

N,N-Dimethyl leucine (DiLeu)

The N,N-Dimethyl leucine, also referred to as DiLeu, is an tandem mass tag reagent which is isobaric and has reporter ions of isotope-encoded dimethylated leucine [[PMID:20218596?](#)]. Each incorporated

label produces a 145.1 Da mass shift. A maximum of four samples can be simultaneously analyzed using DiLeu at a highly reduced cost. MS/MS analysis shows intense reporter ions i.e., dimethylated leucine a1 ions at 115, 116, 117, and 118 m/z. The DiLeu tag labeling efficiency is similar to that of the iTRAQ. Although, DiLeu-labeled peptides offer increased confidence for identification of peptides and more reliable quantification as they undergo better fragmentation thus generating higher reporter ion intensities [PMID:20218596?].

Deuterium isobaric Amine Reactive Tag (DiART)

DiART is an isobaric tagging method used in quantitative proteomics [PMID:22404494?, PMID:20715779?]. The reporter group in DiART tags is a N,N'-dimethyl leucine reporter group with a mass to charge range of 114–119. DiART reagents can label a maximum of six samples and further analyzed by MS. The isotope purity of DiART reagents is very high hence correction of isotopic impurities is not needed at the time of data analysis [PMID:20019052?]. The performances of DiART including the mechanism of fragmentation, the number of proteins identified and the quantification accuracy are similar to iTRAQ. Irrespective of the sequence of the peptide, reporter ions of high-intensity are produced by DiART tags in comparison to those with iTRAQ and thus, DiART labeling can be used to quantify more peptides as well as those with lower abundance, and with reliable results [PMID:22404494?]. DiART serves as a cheaper alternative to TMT and iTRAQ while also having a comparable labeling efficiency. It has been observed that these tags are useful in labeling huge protein quantities from cell lysates before TiO₂ enrichment in quantitative phosphoproteomics studies [PMID:24129742?].

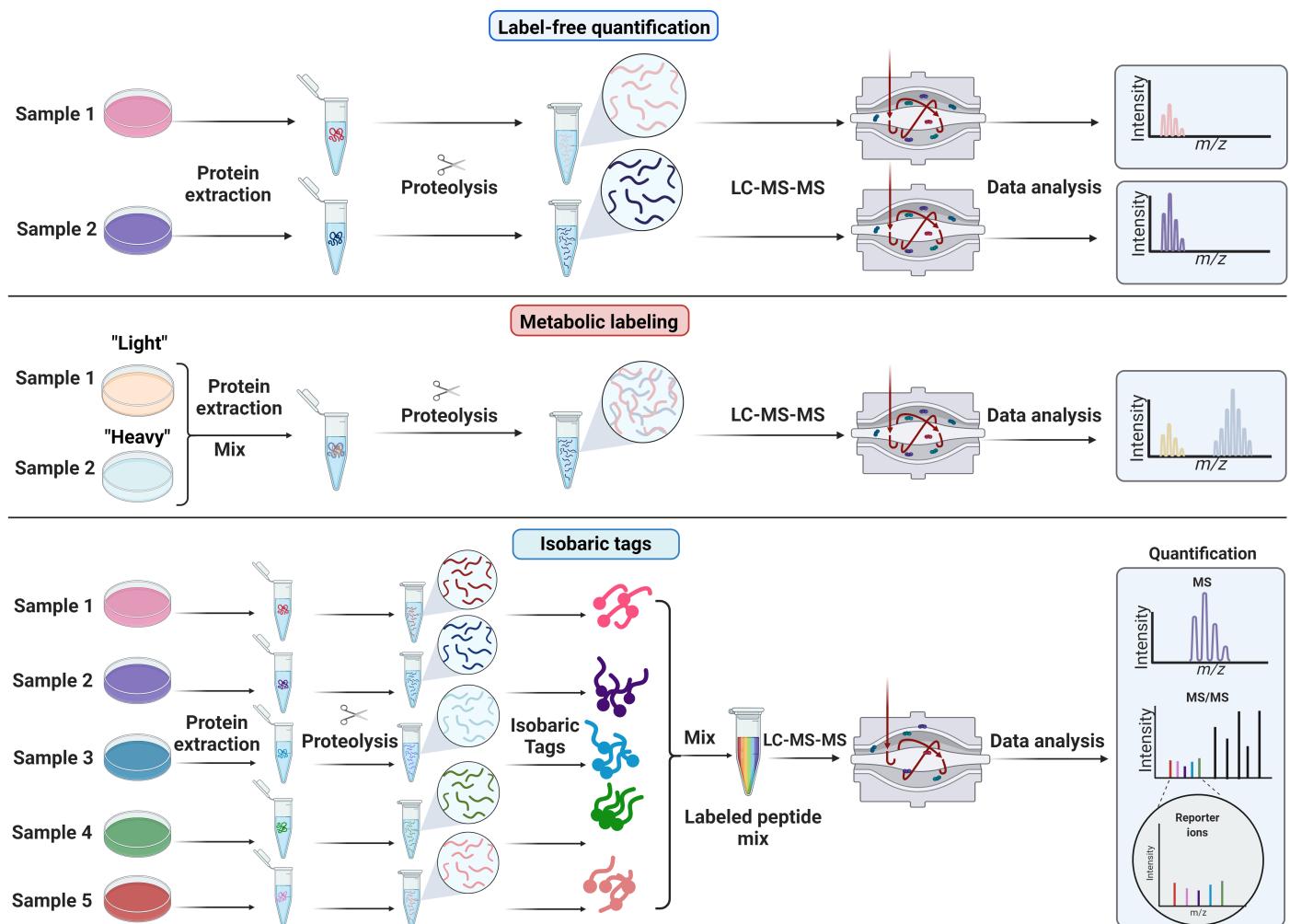


Figure 2: Quantitative strategies commonly used in proteomics. A) Label-free quantitation. Proteins are extracted from samples, enzymatically hydrolyzed into peptides and analyzed by mass spectrometry. Chromatographic peak areas from peptides are compared across samples that are analyzed sequentially. B) Metabolic labelling. Stable isotope labeling with amino acids in cell culture (SILAC) is based on feeding cells stable isotope labeled amino acids ("light" or

"heavy"). Samples grown with heavy or light amino acids are mixed before cell lysis. The relative intensities of the heavy and light peptide are used to compute protein changes between samples. C) Isobaric or chemical labelling. Proteins are isolated separately from samples, enzymatically hydrolyzed into peptides, and then chemically tagged with isobaric stable isotope labels. These isobaric tags produce unique reporter mass-to-charge (m/z) signals that are produced upon fragmentation with MS/MS. Peptide fragment ions are used to identify peptides, and the relative reporter ion signals are used for quantification.

Peptide/Protein Enrichment and Depletion

Protein enrichment (e.g. for protein-protein interactions)

- colP
- APEX
- biOID
- bioplex

Peptide enrichment

- Glycosylation
- SISCAPA

Phosphoproteomics

Protein phosphorylation is when a negatively charged phosphate group (PO_4^-) is added on the side-chain hydroxyl groups of (Ser, S), threonine (Thr, T), and tyrosine (Tyr, Y), residues on target proteins. Protein kinases catalyze the transfer of PO_4^- group from ATP to the nucleophile (OH) group of serine (S), threonine (T), and tyrosine (Y) residues, while protein phosphatases catalyze the removal of PO_4^- . Phosphorylation changes the charge of a protein, often altering protein conformation and therefore function [[PMID:26473910?](#)]. Protein phosphorylation is one of the major post-translation modifications (PTM) that alters the stability, subcellular location, enzymatic activity complex formation, degradation of protein, and cell signalling of protein with a diverse role in cells [[PMID:35227377?](#)]. Phosphorylation can regulate almost all cellular processes, including metabolism, growth, division, differentiation, apoptosis, and signal transduction pathways [[PMID:28656226?](#)]. Rapid changes in protein phosphorylation are associated with several diseases [[PMID:11589691?](#)].

Several methods are used to characterize phosphorylation using modification-specific enrichment techniques combined with advanced MS/MS methods and computational data analysis [[PMID:19743430?](#)]. MS-based phosphoproteomics tools are pivotal for the comprehensive study for the structural and dynamics of cellular signalling networks [[PMID:32676006?](#)], but there are many challenges [[PMID:24890697](#)]. For example, phosphopeptides are low stoichiometry compared to non-phosphorylated peptides, which makes them difficult to identify. Phosphopeptides also exhibit low ionization efficiency [[PMID:28283928?](#)]. To overcome these challenges, it is important to reduce sample complexity to detect large numbers of phosphorylation sites. This is accomplished using enrichment the modified proteins and/or peptides [[PMID:19816929?](#), [PMID:30457327?](#), [PMID:22002794?](#)]. Prefractionation techniques such as strong anionic ion-exchange chromatography (SAX), strong cationic ion-exchange chromatography (SCX), and hydrophilic interaction liquid chromatography (HILIC) are also often useful for reducing sample complexity before enrichment to observe more phosphorylation sites [[106](#)].

As with any proteomics experiment, phosphoproteomics studies require protein extraction, proteolytic enzyme digestion, phosphopeptide enrichment, peptide fractionation, LC-MS/MS, bioinformatics data analysis, and biological function inference. Special consideration is required

during protein extraction where the cell lysis buffer should include phosphatase inhibitors such as sodium orthovanadate, sodium pyrophosphatase and beta-glycerophosphate [[PMID:35011700?](#)].

Enrichment can be done at the protein level before proteolysis. Phosphoprotein enrichment typically involves the use of immobilized metal-affinity chromatography (IMAC) to selectively capture phosphorylated proteins based on their high-affinity binding to metal ions such as Ga(III), Fe(III), Zn(II) and Al(III) [[PMID:16252304?](#)].

Enrichment is more commonly done at the peptide level because there are several advantages over phosphoprotein enrichment. First, peptides have simpler three-dimensional structures than proteins, which makes them easier to separate and analyze. Second, phosphopeptide enrichment is not hindered by small, lipophilic, and very acidic or alkaline proteins [[PMID:22002794?](#)]. Third, prefractionation techniques such as strong anionic ion-exchange chromatography (SAX), strong cationic ion-exchange chromatography (SCX) and hydrophilic interaction liquid chromatography (HILIC) are easier to use for peptide separation than they are for protein separation, and they are more sensitive than 2D-gel electrophoresis, allowing for the identification of less abundant phosphopeptides [[PMID:15966007?](#)]. As a result, phosphopeptide enrichment has more yielded more experimental data than phosphoprotein enrichment [[PMID:17480206?](#)].

Phosphopeptide enrichment often uses titanium dioxide (TiO₂) [[PMID:15858219?](#)] and/or IMAC such as Fe³⁺ coupled to solid-phase materials [[PMID:26584922?](#)]. Often organic acids such as glutamic acid, lactic acid, glycolic acid is added to compete with acidic non-phosphopeptides for binding to the metal-ions.

The use of Fe-IMAC column chromatography allows for the improved phosphopeptide from complex peptide mixtures [[PMID:25394399?](#)]. Compared to other formats like StageTips or batch incubations with TiO₂ or Ti-IMAC beads, Fe-IMAC columns do not suffer from problems with poor binding or elution of phosphopeptides, and the efficiency of enrichment increases linearly with the amount of starting material [[PMID:28188522?](#)].

Multiple IMAC steps can be used in parallel or sequentially to improve phosphopeptide coverage. Lai et al., (2012) showed that the combined use of Fe³⁺-IMAC and Ti(4+)-IMAC chromatography enables complementary identification of more phosphorylation sites than either technique alone [[PMID:22886815?](#)]. A novel phosphopeptide enrichment technique using sequential enrichment with magnetic Fe₃O₄ and TiO₂ particles was developed to detect mono- and multi-phosphorylated peptides [[PMID:20411586?](#)].

General protocol for phosphoproteomics:

- Collect cell line or tissue samples.
- For cell samples, suspend the cell pellet in 2% SDS lysis buffer, heat at 90°C for 5 min, and sonicate.
- For tissue samples, homogenize in liquid nitrogen and add to 4% SDS lysis buffer, then heat at 90°C for 5 min and sonicate.
- Centrifuge the samples at 3,000 rcf to remove insoluble material and remove supernatant to a new tube.
- Add 5 mM Tris (2-carboxyethyl) phosphine (TCEP) and 10 mM chloroacetamide to the samples for alkylation and reduction in the dark at room temperature for 30 minutes.
- Precipitate the protein with acetone to remove SDS.
- Dissolve the protein pellet in 8M Urea with 50 mM TEAB.
- Quantify the protein using a BCA assay.
- Perform proteolysis with trypsin using 1:100 trypsin:substrate (wt:wt). For example, if you have 1 mg of protein, add 10 ug of trypsin.
- Desalt the resulting peptides using Sep-Pak C18 columns or stage tips.

- Subject the fractions to phosphopeptide enrichment using TiO₂/ Fe-IMAC beads.
- Desalt the peptides again using C18 stage tips.
- Perform LC-MS/MS.

Tips for studying phosphorylation:

- Cell lysates should always be prepared using phosphatase inhibitors and samples should be placed on the ice during sonication for protein extraction.
- Increase the amount of starting material of your sample for phosphoenrichment to at least 1 mg of protein or more for optimal results.
- If using anti-phosphorylation antibodies, ensure their specificity is confirmed with other methods.
- Make sure to select a suitable method for the phosphoenrichment that fits the experiment goals.
- TiO₂-based phosphopeptide enrichment methods have different enrichment specificities; selecting non-phosphopeptide excluders such as glutamic acid, lactic acid, glycolic acid, and dihydroxybenzoic acid are the key part of the study [[PMID:35805136?](#)].
- Do not use milk as a blocking agent when western blotting for phosphorylation because milk contains the phosphoprotein casein and can lead to a higher background due to non-specific binding.

Antibody enrichments of modifications

Western blot analysis is used to detect the PTMs in a protein by using antibodies [[PMID:25059473?](#)]. As an extension of this, pan-PTM antibodies have been used to isolate peptides bearing the PTM of interest [[PMID:8633009?](#)]. One benefit of this approach is that peptides are less likely to experience non-specific binding than proteins [[PMID:19743430?](#)]. Initially peptide immunoaffinity precipitation was developed to enrich for phosphotyrosine-containing peptides. This protocol was initially designed to enrich for phosphotyrosine-containing peptides [[PMID:15592455?](#)]. Peptide immunoprecipitation yielded significantly greater coverage of the phosphotyrosine proteome than global phosphorylation enrichment strategies by enriching for a subset of the phosphoproteome. Since then, peptide immunoaffinity precipitation has been used successfully to enrich for peptides with other phosphorylation motifs [[PMID:20736484?](#)] as well as peptides with other modifications such as the diglycyl-lysine residue of ubiquitin modification after trypsin proteolysis [[PMID:21890473?](#)], acetyl-lysine [[PMID:31087289?](#)], arginine methylation [[PMID:24129315?](#)], tyrosine nitration [[PMID:28714690?](#)], and tyrosine phosphorylation [[PMID:15951569?](#)].

The O-linked β-D-N-acetylglucosamine (O-GlcNAc) is found on serine and threonine residues of nucleocytoplasmic proteins and one of PTMs is involved in the occurrence and progression of cancers in multiple systems throughout the body [[PMID:36104770?](#)]. Anti-O-GlcNAc monoclonal antibody enables enrichment from O-GlcNAcylated peptides of cells and tissues. These antibodies have high sensitivity and specificity toward O-GlcNAc-modified peptides and do not identify O-GalNAc or GlcNAc in extended glycans [[PMID:34678516?](#)].

Abundant protein depletion (Blood samples)

The range abundances of proteins in the blood/plasma proteome exceeds 10 orders of magnitude. Due to this wide dynamic range, detection of proteins with medium and low abundance by proteomic analyses is difficult [[PMID:20677825?](#)]. Identifying protein biomarkers from biological samples such as blood is often obstructed by proteins present at higher concentrations. The removal of these high abundant proteins enables the detection of less abundant and unique proteins. The ability to deplete abundant proteins with specificity, reproducibility, and selectivity is extremely important in proteomic studies [[PMID:16052628?](#)].

The following are some of the methods used for abundant protein depletion:

Dye-ligand depletion:

This method is used for the depletion of serum albumin based on the interaction between albumin and dyes like Cibacron Blue (CB) through electrostatic force, hydrogen bonding and hydrophobic interactions. The method is relatively low cost, widely available, robust and has high binding capacity. However, it lacks specificity and has varying efficiency [[PMID:11694290?](#), [PMID:24168355?](#)].

Protein-ligand depletion:

This method is used for depletion of immunoglobulins (Ig) based on the interaction between the Fragment crystallizable (Fc) region of these Igs [[PMID:2473373?](#)] and cell wall protein A, G or A/G of *Staphylococcus aureus* and *Streptococcus* spp [[PMID:2938951?](#), [PMID:10805799?](#)]. It is highly selective and has high yield and purity. However, non-specific binding may occur due to co-absorption of other proteins [[PMID:31617391?](#)].

Immunodepletion:

This method is used for depletion of proteins having high abundance in plasma or serum on the basis of the specific interaction of these proteins with their respective antibodies (antigen-antibody interaction) [[PMID:27896769?](#)]. It has high specificity and commercial kits are also readily available, but it is expensive, has limited sample loading and can result in non-specific binding [[PMID:31617391?](#)].

Combinatorial peptide ligand library:

This method is used for partial depletion of major proteins i.e., those with high abundance and for relative enrichment of lower and medium abundant proteins [[PMID:18451796?](#)]. It is based on the interaction with an array of ligands which are essentially peptides of 6 amino acids in length. It is also used for normalization of the global protein abundance [[PMID:25384740?](#)]. However, the drawbacks include non-specific binding as well as loss of proteins due to incomplete elution or inefficient binding [[PMID:31617391?](#)].

Precipitation:

This method of abundant protein depletion works by altering the solubility of proteins using a chemical reagent including inorganic salt solution [[PMID:21963274?](#)], organic solvents [[PMID:25083595?](#)], non-ionic polymer [[PMID:27832179?](#)] and reducing agents [[PMID:19454248?](#)]. It is extremely simple and cost-effective. However, it is less specific with a risk of protein loss, difficulty in protein resolubilization as well as time consuming [[PMID:31617391?](#)].

New technologies:

Newer methods of highly abundant protein depletion are based on the interaction between polymers such as bacterial cellulose nanofibers [[PMID:30219335?](#)], cryogels [[PMID:30999704?](#), [PMID:23668981?](#)] and nanomaterials [108]. These techniques are highly specific, relatively cheap, and very stable. They can also be reused since they have larger binding capacity and less cross-reactivity [[PMID:31617391?](#)].

Peptide Purification and Fractionation

Peptide purification methods

Solid phase extraction (SPE)

Solid phase extraction (SPE) is a common MS-based proteomics technique employed in the sample preparation. In this method, compound isolation is based on chemical and physical properties, which determines the distribution of compounds between a mobile phase (liquid) and a stationary phase (solid). After the molecules bind, washing of the bound compounds is performed and then molecules are made to elute from the stationary phase after replacing the mobile phase with the elution buffer. The material used for SPE is usually discarded after every sample and no gradient is applied for elution (single-step procedure of elution) [[PMID:14697044?](#)]. Thus, using SPE only a specific analyte group gets separated, which depends on the stationary phase. Hence, SPE is primarily used for sample clean-up and for reducing complexity of the sample. For MS-based proteomic analysis, it is largely used to get rid of salts and other contaminants that might lead to ion suppression. The major drawback of this technique is that with SPE only a small fraction of the sample is examined because not all compounds are captured, but only those with binding capabilities same as that of the sorbent. The material for SPE is available in various types, including (micro-) columns, cartridges, plates, micropipette tips, and functionalized magnetic beads (MBs) [[PMID:20606758?](#), [PMID:20099258?](#)]. Reversed-phase is the most widely used material for SPE in proteomic studies for the proteins and peptide fractionation and rarely, ion-exchange material. For the separation of glycosylated proteins and peptides, the preferred material is normal phase such as HILIC [[\[PMID:22665312?\]](#); 20536156]. SPE materials which are less commonly used are silica- or polystyrene-based ones [[\[PMID:17625912?\]](#); 15317408]. The other types of SPE methods are IEX, metal chelation, and affinity-based [[PMID:25692071?](#)].

The basic idea behind the choice of binding and wash versus elution solutions for SPE is that the binding and wash solutions should favor the interaction between the analytes of interest and the solid phase, whereas the elution solution should favor the interaction of the analyte with the liquid phase. For example, with reversed phase SPE, the solid phase is C18 or some other hydrophobic chemistry. Binding of peptides to this solid phase is based on the hydrophobicity of peptides, mostly due to their peptide backbone, but also due to the presence of amino acid side chains like leucine and phenylalanine. To encourage peptides to 'like' the stationary phase more than the liquid phase, the peptides are loaded in aqueous solution. This will enable washing of the hydrophilic contaminants like salts, small polar buffer molecules, and polar denaturants like urea. After washing the bound peptides, they can be eluted by switching the liquid phase to something hydrophobic, which allows the peptides to partition more into the liquid phase and elute from the solid phase.

Solid Phase Extraction (SPE)

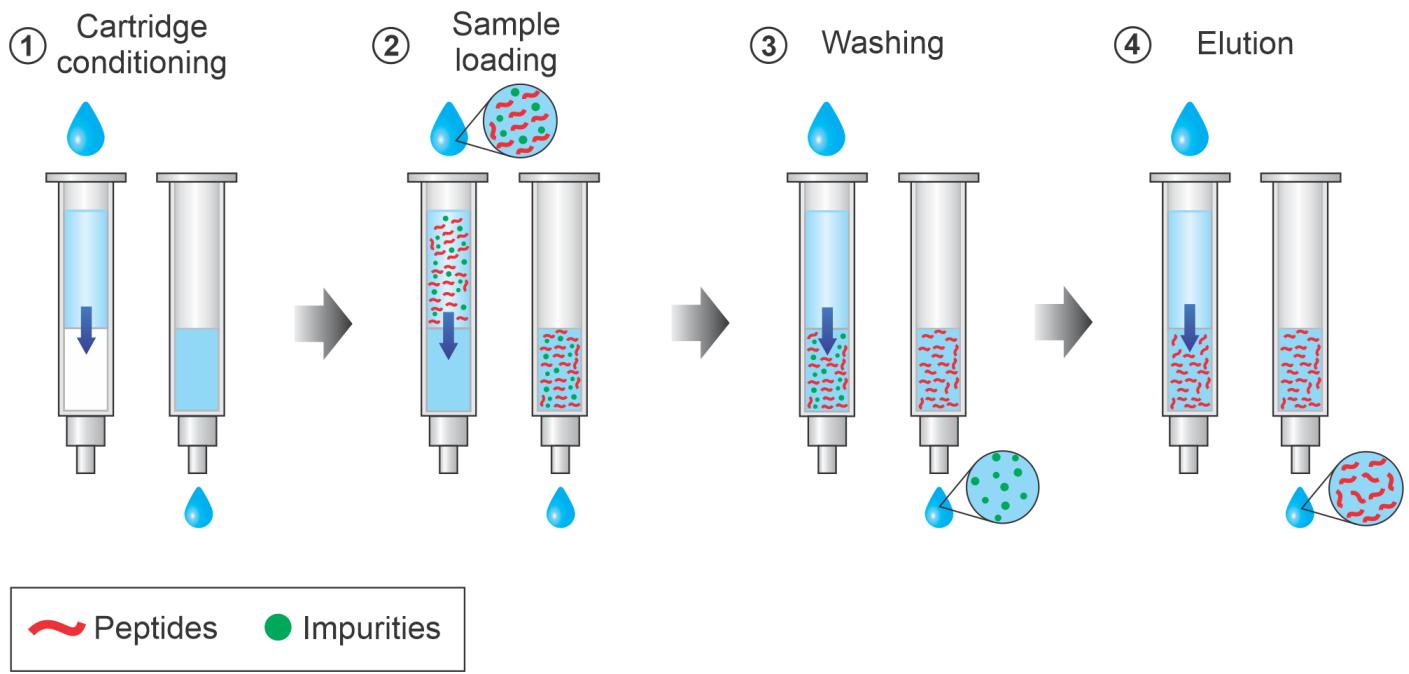


Figure 3: Solid phase extraction (SPE) steps. SPE is a sample preparation technique that uses a solid adsorbent contained most commonly in a cartridge device to selectively adsorb certain molecules from solution. The first step is the conditioning of the cartridge which involves wetting the adsorbent to solvate its functional groups and filling the void spaces with solvent thereby removing any air in the column. This is necessary to produce a suitable environment for adsorption and thus ensure reproducible interaction with the analytes. After conditioning, the sample is loaded in the cartridge. This can be performed with the aid of positive or negative pressure to ensure a constant flow rate. In this step molecules bind the adsorbent and interferences pass through. Next, the column is washed with the mobile phase to eliminate the contaminants while ensuring the analyte remains bound. Finally, peptides are eluted in an appropriate buffer solution with polarity or charge that competes with interaction with the solid phase.

Specific Types of peptide purification

1. Reverse phase including tips and cartridges
2. stage tips
3. in stage tip (iST)
4. SP2, SP3
5. s traps

Peptide fractionation methods

The number of peptides produced from proteolysis of the whole proteome is immense. Thus, after peptides are cleaned from interferences, they are often fractionated into subsets to enable increased proteome coverage. The characterization of the whole proteome is expected from higher order organisms, and with rising interest in post-translational modifications, an elaborate coverage of protein sequence is required. There are different methods for peptide fractionation as follows:

Ion-exchange chromatography (IEC)

This method involves the separation based on contrasting electric charge [[PMID:27868236?](#)]. In this approach, the mechanism of analyte retention is based on the principle of electrostatic attraction between the sample and the stationary phase functional groups (FGs), having opposite charges. IEC is classified into two types: cation-exchange and anion-exchange chromatography. In cation-exchange chromatography, at an acidic pH, the negatively charged functional groups such as sulfates are

attracted to positively charged peptides, whereas, in anion-exchange chromatography, positively charged FGs such as quaternary ammoniums are attracted to peptides with negative charge at an alkaline pH. These techniques are further classified into: strong (cation [SCX] and anion [SAX] exchange), and weak exchangers (cation [WCX] and anion [WAX] exchange), based on the type of FG attached [[PMID:35777803?](#)]. These functional groups are most commonly supported in resins made up of silica and synthetic polymers, however, some inorganic materials are sometimes used [[PMID:27868236?](#)]. In the IEC method, peptide elution is performed using a mobile phase with higher ionic strength, to ensure peptide partition into the liquid phase. SCX along with a salt gradient/plug is a routinely used proteomics technique. In the SCX method, peptides are resolved according to their net charge, in which the peptide with the lowest positive charge is eluted first. Increasing the salt concentration decreases the peptide retention time due to competition with the electrostatic interactions between the peptides and the solid phase. However, SCX resolution is limited compared to reversed phase chromatography and will thus limit the suitability of this technique for complex mixtures [[PMID:15672457?](#)].

Reversed-phase chromatography (RPLC)

Reversed-phase chromatography is the most widely commonly used chromatographic technique which separates molecules in solution having neutral pH based on their hydrophobicity. The separation occurs on the basis of the partition coefficient of analytes between the mobile phase and the hydrophobic stationary phase. Highly polar peptides elute before the ones having less polarity because of the strong interaction with the hydrophobic functional groups forming a layer similar to a liquid around the silica resin [[PMID:20973639?](#)]. RPLC has been widely used in separation of peptides because of its compatibility with gradient elution and aqueous samples and its retention mechanism, which modulates separation owing to changes in the properties like pH, additives and organic modifier [[PMID:20031138?](#)]. Numerous factors influence the capacity of chromatographic peaks, such as temperature, column length, stationary phase, particle size, mobile-phase ion-pairing reagent, mobile-phase modifier and gradient slope [[PMID:16224963?](#)]. Usually online RPLC is done at acidic pH to ensure peptide ionization, but it can be paired with offline high pH RPLC and multiple fraction concatenation to produce orthogonal separation due to altered ionization of amino acids changing peptide hydrophobicity [[PMID:22462785?](#)].

Hydrophilic interaction liquid chromatography (HILIC)

HILIC is similar in its principle to normal-phase chromatography. It is used for the separation of hydrophilic peptides and polar analytes [[PMID:21879300?](#)]. This separation is achieved by a stationary phase that is hydrophilic in nature, for example: cyano-, diol-, amino- bonded phases [[PMID:18428181?](#)], and an organic and hydrophobic mobile phase [[PMID:18264818?](#)]. The elution of bonded peptides occurs by increasing the mobile phase polarity in a reversed elution order as compared to RPLC [[PMID:18264818?](#); 15459207]. Thus, the peptides with less polarity elute before the more polar peptides. HILIC can also be used for enrichment and targeted proteomic analysis of PTMs, such as glycosylation, N-acetylation and phosphorylation, which increase the polarity of peptides and therefore also their retention on HILIC [[PMID:20973639?](#)].

Isoelectric focusing (IEF)

IEF is a type of high-resolution (HR) technique of electrophoresis used for the separation as well as concentration of peptides that are amphoteric in nature on the basis of their isoelectric point (pI) using a solution without buffer consisting of either carrier ampholytes or a gel with immobilized pH gradient (IPG). After IEF separation, the separated amphoteric peptides in the liquid phase are recovered for further analysis by RPLC-MS/MS [[PMID:16849286?](#)]. Along with being a technique with improved resolution and capacity, for separation of peptides, IEF provides with additional information

on physicochemical properties of the peptides, for example: peptide iso electric point (pl) which acts as a tool for validation and filtration for identifying MS/MS peptide sequence during the step of database search [[PMID:18851748?](#)]. The IEF system is not only used for increasing the coverage of proteome but also in quantitative label-free [[PMID:17708596?](#)] and stable isobaric labeling experiments [[PMID:18851748?](#)]. IEF and gel-based separations have fallen out of favor in the last decade due to improvements in liquid chromatography.

Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC)

ERLIC is a method based on use of a weak anion exchange column operated at low pH with high organic solvent enabling isocratic elution [[109](#)]. Acidic peptides are retained by electrostatic interaction, basic and neutral peptides are retained through hydrophilic interaction made favorable by high organic solvent. This improves retention of acidic peptides and reduces retention of basic peptides compared to normal HILIC [[110](#)].

Types of Mass Spectrometers used for Proteomics

1. QQQ
2. Q-TOF
3. Q-Orbitrap
4. LTQ-Orbitrap
5. TOF/TOF
6. FT-ICR
7. types of ion mobility

- SLIM
- FAIMS
- traveling wave
- tims

Peptide Ionization

Until the early 1990s, peptides analysis by mass spectrometry was challenging. Hard ionization techniques in use at the time, like fast atom bombardment, were not directly applicable to peptides without destroying or breaking them. The soft ionization techniques however, revolutionized the proteomics field and it became possible to routinely ionize and analyze peptides using MALDI and ESI techniques at high-throughput scale. These two techniques were so impactful that the 2002 Nobel Prize in Chemistry was co-awarded to John Fenn (ESI) and Koichi Tanaka (MALDI) “for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules” [111].

MALDI

The term, Matrix-assisted LASER desorption ionization (MALDI), was coined by Hillenkamp and Karas in 1985[112]. Karas and Hillenkamp discovered the MALDI technique first, although a similar ionization method was shown by Koichi Tanaka in 1988 [2]. A few months later, Karas and Hillenkamp also demonstrated MALDI applied to protein ionization [113]. It also created a controversy that the widely used method of MALDI from these two people had been overlooked, and the Nobel prize was awarded to Tanaka, whose system was rarely used[114].

MALDI first requires the peptide sample to be co-crystallized with a matrix molecule, which is usually a volatile, low molecular-weight, organic aromatic compound. Some examples of such compounds are cyano-hydroxycinnamic acid, dihydrobenzic acid, sinapinic acid, alpha-hydroxycinnamic acid, ferulic acid etc [115/]. Subsequently, the analyte is placed in a vacuum chamber in which it is irradiated with a LASER, usually at 337nm [116]. This laser energy is absorbed by the matrix, which then transfers that energy along with its free protons to the co-crystallized peptides without significantly breaking them. The matrix and co-crystallized sample generate plumes, and the volatile matrix imparts its protons to the peptides as it gets ionized first. The weak acidic conditions used as well as the acidic nature of the matrix allows easy exchange of protons for the peptides to get ionized and fly under the electrical field in the mass spectrometer. These ionized peptides generally form the metastable ions, most of them will fragment quickly [117]. However, it can take several milliseconds and the mass spectrometry analysis can be performed before this time. Peptides ionized by MALDI almost always take up a single charge and thus observed and detected as $[M+H]^+$ species.

MALDI Mechanism

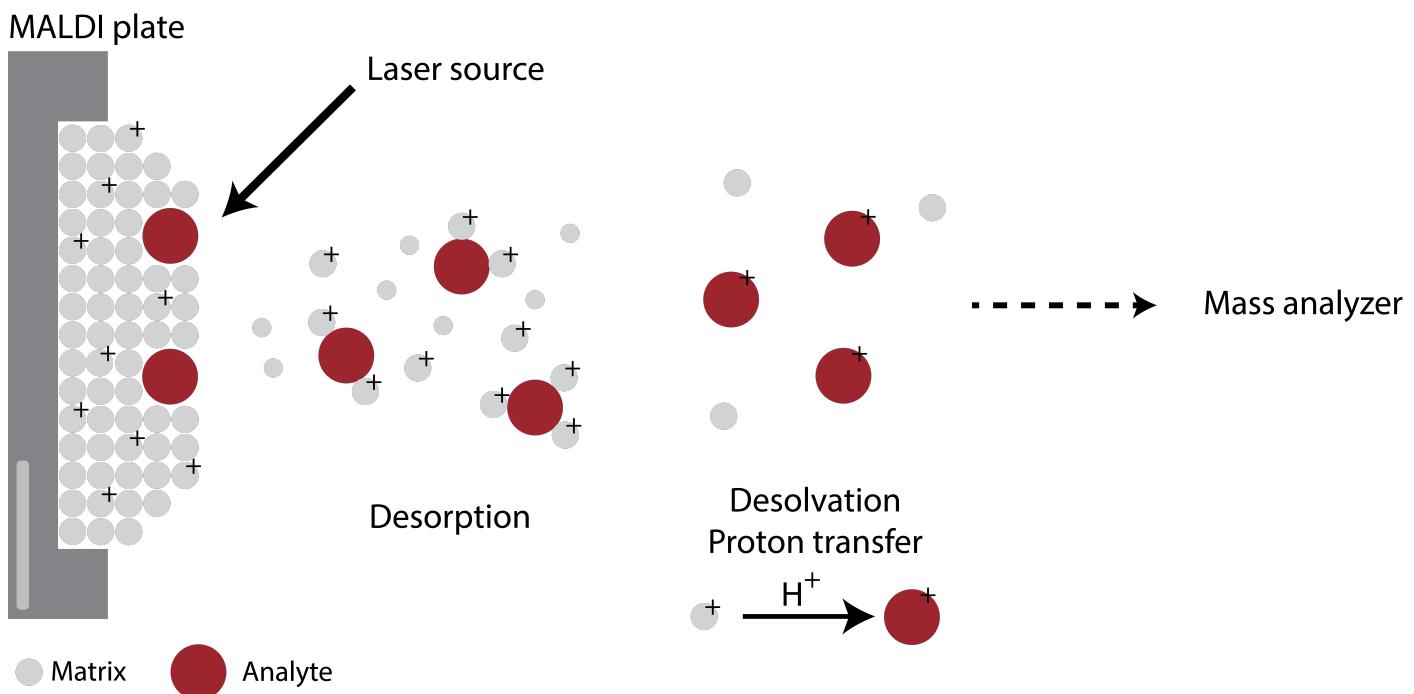


Figure 1: MALDI The analyte-matrix mixture is irradiated by a laser source, leading to ablation. Desorption and proton transfer ionize the analyte molecules that can then be accelerated into a mass spectrometer.

Electrospray Ionization

ESI was first applied to peptides by John Fenn and coworkers in 1989 [1]. Concepts related to electrospray ionization (ESI) were published at least as early as 1882, when Lord Rayleigh described the number of charges that could assemble on the surface of a droplet [118]. ESI is usually coupled with reverse-phase liquid-chromatography of peptides directly interfaced to a mass spectrometer. A high voltage (~ 2 kV) is applied between the spray needle and the mass spectrometer. As solvent exits the needle, it forms droplets that take on charge at the surface, and through a debated mechanism, those charges are imparted to peptide ions. The liquid phase is generally kept acidic to help impart protons easily to the analytes.

Tryptic peptides ionized by ESI usually carry one charge on the side chain of their c-terminal residue (Arg or Lys) and one charge at their n-terminal amine. Peptides can have more than one charge if they have a longer peptide backbone, have histidine residues, or have missed cleavages leaving extra Arg and Lys. In most cases, peptides ionized by ESI are observed at more than one charge state. Evidence suggests that the distribution of peptide charge states can be manipulated through chemical additives [[PMID:22610994?](#)].

Electrospray Mechanism

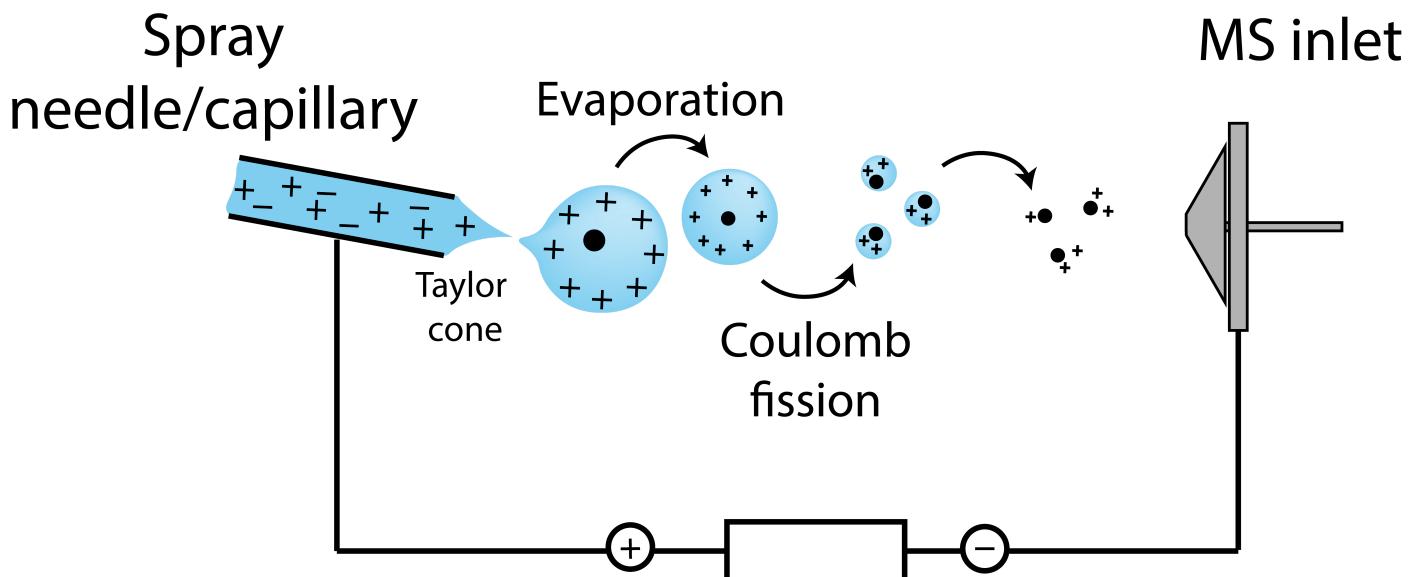


Figure 1: Electrospray Ionization Charged droplets are formed, their size is reduced due to evaporation until charge repulsion leads to Coulomb fission and results in charged analyte molecules.

The main goal of ESI is the production of gas-phase ions from electrolyte ions in solution. During the process of ionization, the solution emerging from the electrospray needle or capillary is distorted into a Taylor cone and charged droplets are formed. The charged droplets subsequently decrease in size due to solvent evaporation. As the droplets shrink, the charge density and Coulombic repulsion increase. This process destabilizes the droplets until the repulsion between the charges is higher than the surface tension and they fission (Coulomb explosion) [PMID:19551695?] [PMID:23134552?]. Typical bottom-up proteomics experiments make use of acidic analyte solutions which leads to the formation of positively charged analyte molecules due to an excess presence of protons.

Data Acquisition

Hybrid mass spectrometers used for modern proteome analysis offer the flexibility to collect data in many different ways. Data acquisition strategies differ in the sequence of precursor scans and fragment ion scans, and in how analytes are chosen for MS/MS. Constant innovation to develop better data collection methods improves our view of the proteome, but many method options may confuse newcomers. This section provides an overview of the general classes of data collection methods.

Data acquisition strategies for proteomics fall into one of two groups.

1. Data dependent acquisition (DDA), in which the exact scan sequence in each analysis depends on the data that the mass spectrometer observes.
2. Data independent acquisition (DIA), in which the exact scan sequence in each analysis DOES NOT depend on the data; the collected scans are the same whether you inject yeast peptides, human peptides, or a solvent blank.

DDA and DIA can both be further subdivided into targeted and untargeted methods.

DDA

In most cases, the peptide masses that will be observed are not known before doing the experiment. Data collection methods must account for this. DDA was invented in the early 1990s, which enabled collecting MS/MS spectra for observed peptides as they eluting from the LC column [\[119,120, PMID:24203425?\]](#).

Untargeted DDA

A common method currently used in modern proteomics is untargeted DDA. The MS collects precursor (MS1) scans iteratively until precursor mass envelopes meeting certain criteria are detected. Criteria for selection are usually specific charge states and a minimum signal intensity. When those ions meet these criteria, the MS selects those masses for fragmentation.

Because ions are selected as they are observed, repeated DDA of the same sample will produce a different set of identifications. This stochasticity is the main drawback of DDA. To ameliorate this issue, often strategies are used to transfer identifications between multiple sample analyses. This transfer of IDs across runs is known as “match between runs”, which was originally made famous by the processing software MaxQuant [\[121,122\]](#). There are several other similar tools and strategies, including the accurate mass and time approach [\[123\]](#), Q-MEND [\[124\]](#), IDEAL-Q [\[125\]](#) and superHIRN [\[126\]](#). More recent work has introduced statistical assessment of MBR methods using a two-proteome model [\[127\]](#). Statistically controlled MBR is currently available in the IonQuant tool [\[128\]](#).

Because DDA is required for quantification of proteins using isobaric tags like TMT, this stochasticity of DDA limits the ability to compare quantities across batches. For example, if you have 30 samples, you can use two sets of the 16-plex kit to label 15 samples in each set with one channel labeled by a pooled sample to enable comparison across the groups. When you collect DDA data from each of those sets, each set will have MS/MS data from an overlapping but different set of peptides. If one set has MS/MS from a peptide but the other set does not, then that peptide cannot be quantified in the whole sample group. This limits the number of quantified proteins in large TMT experiments with multiple batches.

Targeted DDA

Targeted DDA is not common in modern proteomics. In targeted DDA, in addition to general criteria like a minimum intensity and a certain charge state, the mass spectrometer looks for specific masses. These masses might be previously observed signals that were previously missed by MS/MS [129,130]. In these studies, the sample is first analyzed by LC-MS to detect precursor ion features with some software, and then subsequent analyses target those masses for fragmentation with inclusion lists until they are all fragmented. This was shown to increase proteome coverage.

DDA methods for modifications

There are also DDA methods that look for specific fragment or neutral loss ions in the resulting spectra. For example, when linear ion traps were the main proteomics workhorses, CID analysis of phosphopeptides would result in predominantly neutral loss of the phosphate with limited sequence ion information. To gain sequence ions in these experiments, instruments could be set to isolate a loss of 98 Thomspons for MS3 [131,132]. The newer collisional dissociation technique HCD significantly improves the detection of peptide fragments with the phosphorylation intact on fragment ions, and thus, this neutral loss scanning technique is no longer common.

A similar strategy was introduced for N-linked glycopeptides [133]. Collisional dissociation of glycosylated peptides produces oxonium ions at 204.09 (HexNAc) or 366.14 (HexHexNAc). In oxonium ions from the glycosylation were detected in the top 20 fragments of the HCD spectra, then an ETD scan was triggered. This ETD scan provides information about the peptide sequence, while the original HCD scan provides glycan structure information.

DIA

The simplest method to operate a mass spectrometer is to have predefined scans that are collected for each sample analysis. This is DIA; the scans that are collected do not depend on the data that the instrument observes. Thus, the scan sequence is the same every time. Although simple in terms of data collection, when the scan sequence includes MS/MS, sophisticated software is required to analyze the data. Like DDA, DIA can also be either targeted or untargeted [134]: The two targeted DIA methods are selected reaction monitoring (SRM) or multiple reaction monitoring (MRM), and untargeted DIA (uDIA) is often referred to simply as "DIA" or SWATH (Figure 4).

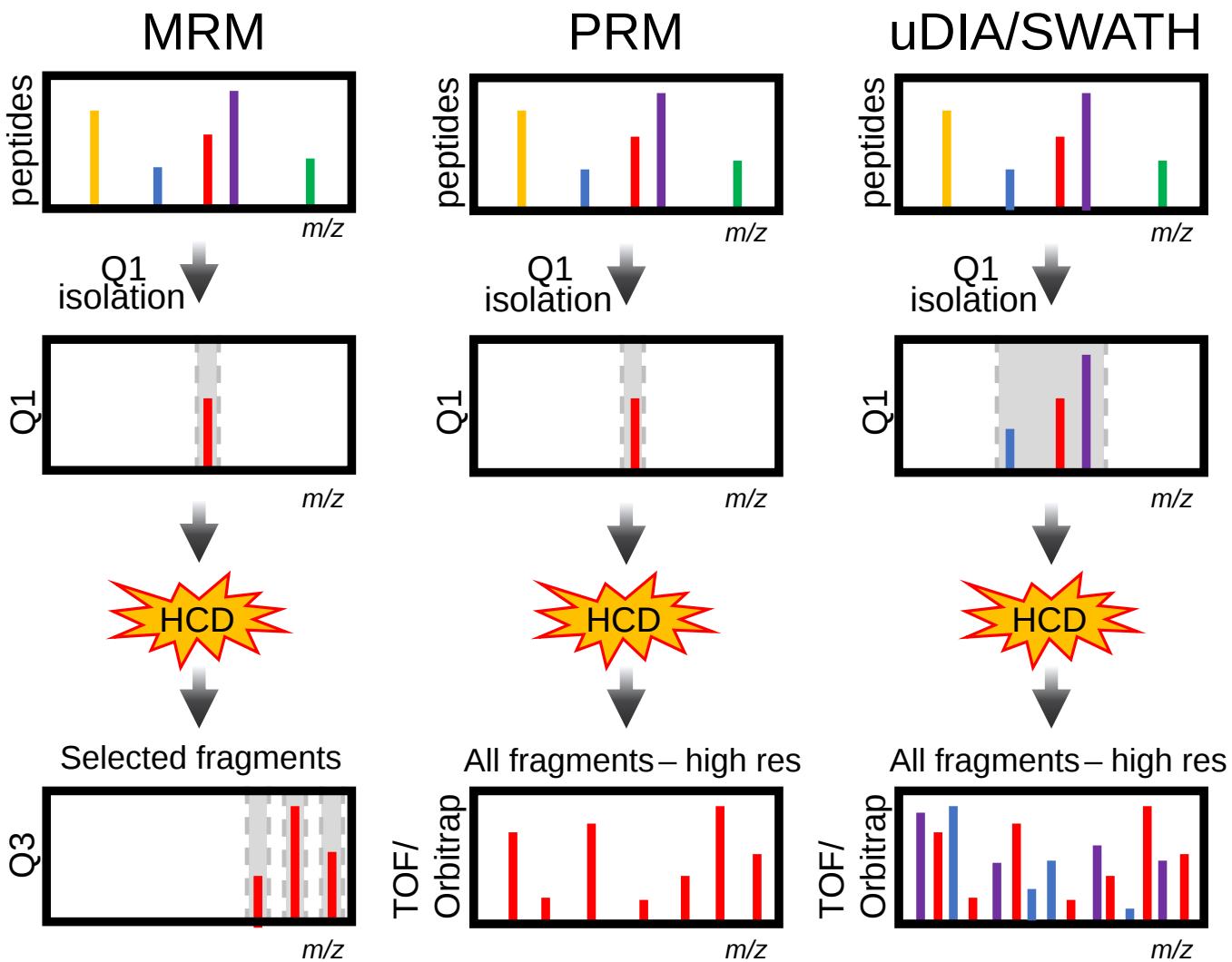


Figure 4: Types of DIA. A) SRM/MRM. Peptides are ionized by ESI and although there are many peptides entering the mass spectrometer at any time, the first quadrupole (Q1) isolates one mass, which is then fragmented by HCD. Fragment masses from the peptide are then selected in the third quadrupole (Q3). This leads to very low noise and high sensitivity. B) PRM. Like MRM, peptides are selected in the first quadrupole, but this analysis is done on a high resolution instrument like an Orbitrap or TOF. Selectivity is gained by exploiting the high mass accuracy and resolution to monitor multiple fragment ions. C) uDIA/SWATH. Like MRM and PRM, peptides are isolated with Q1, but in this case a much wider isolation window is used. This usually results in co-isolation of many peptides simultaneously. Fragments from many peptides are measured with high resolution and high mass accuracy. Special software is used to get peptide identities and quantities from the fragment ions.

Targeted DIA

The first type of targeted DIA is called SRM or MRM [135]. The popularity of this method in the literature peaked in 2014, with just under 1,500 documents on pubmed that year resulting from a search for "MRM". In this strategy, the QQQ MS is set so that the first quadrupole selects the precursor mass of the peptide(s) of interest, the second quadrupole fragments the peptide, and the third quadrupole monitors the product of specific fragments from that peptide. This strategy is very sensitive and has the benefit of very low noise. The fragments monitored in Q3 are chosen such that it is unlikely these fragments could arise from another peptide. Usually at least a few transitions are monitored for each peptide in order to get multiple measures for that peptide.

An early example of MRM applied to quantify c-reactive protein was in 2004 [136]. Around the same time, SRM was combined with antibody enrichment of peptides from target proteins [137]. This approach was popular for analysis of plasma proteins [138]. These early examples led to many more studies that used QQQ MS instruments to get accurate quantitation of many proteins in one injection.

[139,140]. Scheduling MRM measurement when chromatography is stable additionally enabled better utilization of instrument duty cycle and therefore monitoring of more peptides per injection [141]. Efforts even developed libraries of transitions that allow quantification of any protein in model organisms [142].

Another similar targeted DIA method is called parallel reaction monitoring (PRM) [143]. Instead of using a QQQ instrument, PRM uses a hybrid MS with a quadrupole and a high resolution mass analyzer, such as an Q-TOF or Q-Exactive. The idea is that instead of monitoring specific fragments in Q3, the high mass accuracy can be used to filter peptide fragments for high selectivity and accurate quantification. Studies have found that PRM and MRM/SRM have comparable dynamic range and linearity [144].

Untargeted DIA

There were many implementations of uDIA over the years, starting in 2003 by Purvine et al from the Goodlett lab [145]. In this first work they demonstrated uDIA using a Q-TOF with in source fragmentation, and showed that extracted ion chromatograms of precursor and fragment ions matched in shape suggesting that this could be used to identify and quantify peptides. The following year, Venable et al from the Yates lab introduced uDIA with an ion trap [146]. Subsequent methods include MSE [147], PAcIFIC [148], all ions fragmentation (AIF) [149]. Computational methods were also developed to automate interpretation of this data, such as DeMux [149], XDIA [150], and ETISEQ [151].

The paper that is often cited for uDIA that led to widespread adoption was by Gillet et al. from the Aebersold group in 2012 [152]. In this paper they branded the idea as SWATH. Widespread adoption may have been facilitated by the co-marketing of this idea by ABSciex as a proteomics solution on their new 5600 Q-TOF (called “tripleTOF” despite containing only one TOF, likely a portmanteau of “triple quadrupole” and “Q-TOF”). Importantly, in the Gillet et al. paper the authors described a computational method to extract information from SWATH where peptides of interest were queried against the data. They also demonstrated the application of SWATH to measure proteomic changes that happen in diauxic shift, and showed that SWATH can reveal modified peptides, in this case a methionine oxidation.

There are also many papers describing uDIA with orbitraps. One early example described combining random isolation windows together and then demultiplexing the chimeric spectra [153]. In another landmark paper, over 6,000 proteins were identified from mouse tissue by at least 2 peptides [154]. In 2018, the new model orbitrap at that time (HF-X) enabled identification of nearly 6,000 human proteins in only 30 minutes. Currently orbitraps have all but replaced the Sciex Q-TOFs for DIA data collection.

A new direction in uDIA is the addition of ion separation by ion mobility. This has appeared in two forms. On the timsTOF, diaPASEF makes use of the trapped ion mobility to increase speed and sensitivity of analysis [155]. On the orbitrap, the combination of FAIMS and DIA has enabled the identification of over 10,000 proteins from one sample, which is a major milestone [156].

Analysis of Raw Data

The goal of basic data analysis is to convert raw spectral data into identities and quantities of peptides and proteins that can be used for biologically-focused analysis. This step may often include measures of quality control, cross-run data normalization, quantification on different levels (precursor, peptide, protein), protein inference, PTM (post translational modification) localization and also first steps of data analysis, such as statistical hypothesis tests.

In typical bottom-up proteomics experiments, proteins are digested into peptides and further analyzed with LC-MS/MS systems. Peptides can have different PTMs and ionize differently depending on their length and amino acid distributions. Therefore, mass spectrometers often record different charge and modification states of one single peptide. The entity that is recorded on a mass spectrometer is usually referred to as a precursor ion (peptide with its modification and charge state). This precursor ion is fragmented and the precursor or peptide sequences are obtained through spectral matching. The quantity of a precursor is estimated with various methods. The measured precursor quantities are combined to generate a peptide quantity. Peptides are also often combined into a protein group through protein inference, which combines multiple peptide identifications into a single protein identification [[PMID:16009968?](#)] [157]. Protein inference is still a challenge in bottom-up proteomics.

Due to the inherent differences in the data structures of DDA and DIA measurements, there exist different types of software that can facilitate the steps mentioned above. The existing software for DDA and DIA analysis can be further divided into freeware and non-freeware:

DDA freeware

Name	Publication	Website
MaxQuant	Cox and Mann, 2008[158]	MaxQuant
MSFragger	Kong et al., 2017[159]	MSFragger
Mascot	Perkins et al., 1999[PMID:10612281?]	Mascot
MS-GF+	Kim et al., [160]	MS-GF+
X!Tandem	Craig et al., [161 , 162]	GPMDB

DIA freeware:

Name	Publication	Website
MaxDIA	Cox and Mann, 2008[158]	MaxQuant
Skyline	MacLean et al., 2010[163]	Skyline
DIA-NN	Demichev et al., 2019[164]	DIA-NN

Targeted proteomics freeware:

Name	Publication	Website
Skyline	MacLean et al., 2010[163]	Skyline

DDA non-freeware:

Name	Publication	Website
ProteomeDiscoverer		ProteomeDiscoverer
Mascot	Perkins et al., 1999 [PMID:10612281?]	Mascot
Spectromine		Spectromine
PEAKS	Tran et al., 2018 [165]	PEAKS

DIA non-freeware:

Name	Publication	Website
Spectronaut	Bruderer et al., 2015 [166]	Spectronaut
PEAKS	Tran et al., 2018 [165]	PEAKS

Data Summary and Interpretation

Name	Publication	Website
Peptide Shaker	Vaudel et al., 2015 [167,168]	PeptideShaker , Peptide Shaker Online

Analysis of DDA data

DDA data analysis either directly uses the vendor proprietary data format directly with a proprietary search engine like Mascot, Sequest (through Proteome Discoverer), Paragon (through Protein Pilot), or it can be processed through one of the many freely available search engines or pipelines, for example, MaxQuant, MSGF+, X!Tandem, Morpheus, MSFragger, and OMSSA. Tables 1 and 4 give weblinks and citations for these software tools. For analysis with freeware, raw data is converted to either text-based MGF (mascot generic format) or into a standard open XML format like mzML [[169](#)][[\[PMID:20013381?\]](#)][[170](#)]. The appropriate FASTA file containing proteins predicted from that organism's genome is chosen as a reference database to search the experimental spectra. All search parameters like peptide and fragment mass errors (i.e. MS1 and MS2 tolerances), enzyme specificity, number of missed cleavages, chemical artefacts (fixed modifications) and potential biological modifications (variable/dynamic modifications) are specified before executing the search. The search algorithm scores each query spectrum against its possible peptide matches [[171](#)]. A spectrum and its best scoring candidate peptide are called a peptide spectrum match (PSM). The scores reflect a *goodness-of-fit* between an experimental spectrum and a theoretical one and do not necessarily depict the correctness of the peptide assignment.

For evaluating the matches, a decoy database is preferred as a null model for peptide matching. A randomized or reversed version of target database is used as a nonparametric null model. The decoy database can be searched separate from the target database (Kall's method)[[172](#)] or it can be combined with the target database before search (Elias and Gygi method)[[PMID:17327847?](#)]. Using either separate method or concatenated database search method, an estimate of false hits can be calculated which is used to estimate the false discovery rate (FDR) [[173](#)]. The FDR denotes the proportion of false hits in the population accepted as true. For Kall's method: the false hits are estimated to be the number of decoys above a given threshold. It is assumed that the number of decoy hits that pass a threshold are the false hits. A similar number of target population may also be false. Therefore, the FDR is calculated as [[174](#)]:

$$FDR = \frac{DecoyPSMs + 1}{TargetPSMs}$$

For Elias and Gygi Method, the target population in which FDR is estimated changes. The target and decoy hits coming from a joint database compete against each other. For any spectrum, either a target or a decoy peptide can be the best hit. It is argued that the joint target-decoy population has decoy hits as confirmed false hits. However, due to the joint database search, the target database may also have equal number of false hits. Thus, the number of false hits is multiplied by two for FDR estimation.

$$FDR = \frac{2 * DecoyPSMs}{Target + DecoyPSMs}$$

Strategies for analysis of DIA data

Targeted proteomics data analysis

Quality control

Quality control should be a central aspect of any mass spectrometry-based study to ensure reproducibility of generated results. There are two types of quality controls that can be conducted for any kind of mass spectrometry experiment. The first one is focused on monitoring the performance of the instruments themselves (e.g. HPLC and mass spectrometer), whereas the second one is focused on your experiments. For further reading, we recommend to take a look at issue 11 on quality control published in the journal *Proteomics* in 2011 [175], especially the review by Köcher *et al.* [176], as well as the review published by Bittremieux *et al.* in 2017 [177].

Instrument Performance

It is generally advisable to monitor instrument performance regularly. Instrument calibrations in regular intervals help ensure that performance is maintained. Often basic calibration and sensitivity can be checked by direct infusion of a standard. During the calibration you can check injection times (for ion trap instruments) and intensity of the ions in the calibration mix.

After ensuring good calibration and signal with the simple calibration mixture, it is advisable to analyze complex samples, such as tryptic digests of whole-cell lysates (e.g. HeLa cells, HEK cells, yeast, etc.) or tryptic digests of purified proteins. The additional check with a complex sample ensures all aspects of the system are working together correctly, especially the liquid chromatography and emitter. These digests should be analyzed after every instrument calibration and periodically between samples when acquiring more extensive batches. Data measured from tryptic digests should be analyzed by the software of your choice and the numbers of identified peptide precursors and proteins can be compared with previous controls for consistency.

Another strategy is to analyze digested purified proteins, which easily enable discovery of retention time shifts and mass accuracy problems. In case you are working with a Thermo mass spectrometer, you can open the acquired .raw file directly either in FreeStyle or in Qual Browser and look for specific m/z values of your peptides. Looking at the intensity of the extracted peaks will help identify sensitivity fluctuations.

Carry-over between different measurements can be identified from blank measurements which are subsequently analyzed with your search software of choice. Blank measurements can be injections of

different buffers, water or the starting conditions of your liquid chromatography. In case of increased detection of carry-over, injections with trifluoroethanol can be performed.

Another factor to take into consideration is the stability of your electrospray. Electrospray stability tends to worsen over time as columns wear, as well as when measuring samples with residual contaminants, such as salts or detergents. You will notice spray instabilities either in the total ion chromatogram (TIC) as thin spikes with short periods of no measured signal or if you install cameras at your ESI source. Suboptimal spray conditions will usually result in droplets forming on the emitter, being released into the mass spectrometer (also referred to as "spitting"). Real-time quality control software (listed in the table below) can help you identify instrument issues right away.

Data Quality Control

Apart from instrument performance, any kind of data analysis should have proper quality control in place to identify problematic measurements and to exclude them if necessary. It is recommended to develop a standardized system for data quality control early on and to keep this consistent over time. Adding indexed retention time (iRT) peptides can help identify and correct gradient and retention time inconsistencies between samples at the data analysis stage. Decoy searches help monitor and control the false-discovery rate. Including common contaminants, such as keratins, in the FASTA files used for searches can help identify sample preparation issues. Other parameters to check in your analysis are the consistency of the number of peptide-spectrum matches, identified peptides and proteins over all samples of your study, as well as your coefficients of variation between your replicates. Before and after data normalization (if normalization is performed) it is good to compare the median intensities of all measurements to identify potential measurement or normalization issues. Precursor charge distributions, missed cleavage numbers, peak width, as well as the number of points per peak are additional parameters that can be checked. In case you are analyzing different conditions, you can perform hierarchical clustering or a principal component analysis to check if your samples cluster as expected.

Quality Control Software

Raw file and real-time analysis

Name	Supported instrument vendors	Website/Download	publication	Note
QuiC	Thermo Scientific, AB SCIEX, Agilent, Bruker, Waters	QuiC		requires Biognosys iRT peptides
AlphaPept	Thermo Scientific, Bruker	AlphaPept	[178]	
RawMeat 2.1	Thermo Scientific	RawMeat		
rawDiag	Thermo Scientific	rawDiag	[179]	
rawrr	Thermo Scientific	rawrr	[180]	
rawBeans	Thermo or mzML	rawBeans	[181]	
SIMPATIQCO	Thermo Scientific	SIMPATIQCO	[182]	

Name	Supported instrument vendors	Website/Download	publication	Note
QC-ART		QC-ART	[183]	
Spray Qc	Thermo Scientific, AB SCIEX, extensible to other instrumentation	SprayQc	[184]	
Metriculator		Metriculator	[185]	
Mass QC		MassQC		
Open MS		OpenMS	[186]	

Search result QC

Name	Website/Download/publication	publication	Note
MSStats	MSStats	[187]	can use output from MaxQuant, Proteome Discoverer, Skyline, Progenesis, Spectronaut
MSStatsQC	MSStatsQC	[188]	
PTXQC	PTXQC	[189]	requires MaxQuant search engine output
protti	protti	[190]	

Statistical hypothesis testing

Databases

What are they and where do you get them?

Protein Database Sources and Types

Many mass spectrometry-based proteomic techniques use search algorithms that require a defined theoretical search space to identify peptide sequences based on precursor mass and fragmentation patterns, which are then used to infer the presence and abundance of a protein. The search space is calculated from the potential proteins in a sample, which includes the proteome (often a single species) and expected contaminants. This is called database searching and the flat file of protein sequences in FASTA format acts as a protein database. In this section, we will describe major resources for proteome FASTA files (protein sequence collections), how to retrieve them, and suggested best practices for preserving FASTA file provenance to improve reproducibility.

In general, FASTA sequence collections can be retrieved from three central clearing houses: UniProt, RefSeq, and Ensembl. These will be discussed separately below as they each have specific design goals, data products, and unique characteristics. It is important to learn the following three points for each resource: the source of the underlying data, canonical versus non-canonical sequences, and how versioning works. These points, along with general best practices, such as using a taxonomic identifier, are essential to understand and communicate search settings used in analyses of proteomic datasets. Finally, it is critical to understand that sequence collections from these three resources are not the same, nor do they offer the same sets of species.

Key terminology may vary between resources, so these terms are defined here. The term “taxon identifier” is used across resources and is based on the NCBI taxonomy database. Every taxonomic node has a number, e.g., *Homo sapiens* (genus species) is 9606 and *Mammalia* (class) is 40674. This can be useful when retrieving and describing protein sequence collections. Another term used is “annotation”, which has different meanings in different contexts. Broadly, a “genome annotation” is the result of an annotation pipeline to predict coding sequences, and often a gene name/symbol if possible. Two examples are MAKER [[PMID:22192575?](#)] and the RefSeq annotation pipeline [[191](#)]. Alternatively, “protein annotation” (or gene annotation) often refers to the annotation of proteins (gene products) using names and ontology (i.e., protein names, gene names/symbols, functional domains, gene ontology, keywords, etc.). Protein annotation is termed “biocuration” and described in detail by UniProt [[192](#)]. Lastly, there are established minimum reporting guidelines for referring to FASTA files established in MIAPE: Mass Spectrometry Informatics that are taxon identifier and number of sequences [[193, PMID:23500130?](#)]. The FASTA file naming suggestions below are not official but are suggested as a best practice.

UniProt

The Universal Protein Resource (UniProt) [[PMID:14681372?](#)], has three different products: UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). The numerous resources and capabilities associated with the UniProt are not explored in this section, but these are well described on UniProt’s website. UniProtKB is the source of proteomes across the Tree of Life and is the resource we will be describing herein. There are broadly two types of proteome sequence collections: Swiss-Prot/TrEMBL and designated proteomes. The Swiss-Prot/TrEMBL type can be understood by discussing how data is integrated into UniProt. Most protein sequences in UniProt are derived from coding sequences submitted to EMBL-Bank, GenBank and DDBJ. These translated sequences are initially imported into TrEMBL database, which is why TrEMBL is also termed “unreviewed”. There are other sources of protein sequences, as described by UniProt

[194]. These include the Protein Data Bank (PDB), direct protein sequencing, sequences derived from the literature, gene prediction (from sources such as Ensembl) or in-house prediction by UniProt itself. Protein sequences can then be manually curated into the Swiss-Prot database using multiple outlined steps (described in detail by UniProt here [195]) and is why Swiss-Prot is also termed “reviewed”. Note that more than one TrEMBL entry may be removed and replaced by a single Swiss-Prot entry during curation. A search of “organism:9606” at UniProtKB will retrieve both the Swiss-Prot/reviewed and TrEMBL/unreviewed sequences for *Homo sapiens*. The entries do not overlap, so users often either use just Swiss-Prot or Swiss-Prot combined with TrEMBL, the latter being the most exhaustive option. With ever-increasing numbers of high-quality genome assemblies processed with robust automated annotation pipelines, TrEMBL entries will contain higher quality protein sequences than in the past. In other words, if a mammal species has 20 000 to 40 000 entries in UniProtKB and many of these are TrEMBL, users should be comfortable using all the protein entries to define their search space (more on this later when discussing proteomes at UniProtKB). Determining the expected size of a well-annotated proteome requires additional knowledge, but tools to answer these questions continue to improve. As more and more genome annotations are generated, the backlog of manual curation continues to increase. However, automated genome annotations are also rapidly improving, blurring the line between Swiss-Prot and TrEMBL utility.

The second type of protein sequence collections available at UniProtKB are designated proteomes, with subclasses of “proteome”, “reference proteome” or “pan-proteome”. As defined by UniProt, a proteome is the set of proteins derived from the annotation of a completely sequenced genome assembly (one proteome per genome assembly). This means that a proteome will include both Swiss-Prot and TrEMBL entries present in a single genome annotation, and that all entries in the proteome can be traced to a single complete genome assembly. This aids in tracking provenance as assemblies change, and metrics of these assemblies are available. These metrics include Benchmarking Universal Single-Copy Ortholog (BUSCO) score, and “Completeness” as Standard, Close Standard or Outlier based on the Complete Proteome Detector (CPD). Given the quality of genome annotation pipelines, using a proteome as a FASTA file for a species is the preferred method of defining search spaces now. Outside of humans, no higher eukaryotic Swiss-Prot sequence collections are complete enough for use in proteomics analyses, but this does not mean that the available Swiss-Prot plus TrEMBL protein sequence collection precludes accurate proteomic data analysis. Lastly, the difference between reference proteome and proteome is used to highlight model organisms or organisms of interest, but not to imply improved quality. UniProt also has support for the concept of “pan proteomes” (consensus proteomes for a closely related set of organisms) but this is mostly used for bacteria (e.g., strains of a given species will share a pan proteome).

When retrieving protein sequence collections as Swiss-Prot/TrEMBL or designated proteomes, there is an option of downloading “FASTA (canonical)” or “FASTA (canonical & isoform)”. The later includes additional manually annotated isoforms for Swiss-Prot sequences. Each Swiss-Prot entry has one canonical sequence chosen by the manual curator. Any additional sequence variants (mostly from alternative slicing) are annotated as differences with respect to the canonical sequence. Specifying “canonical” will select only one protein sequence per Swiss-Prot entry while specifying “canonical & isoforms” will download additional protein sequences by including isoforms for Swiss-Prot entries. Recently, an option to “download one protein sequence per gene (FASTA)” has been added. These FASTA files include Swiss-Prot and TrEMBL sequences to number about 20 000 protein sequences for a wide range of higher eukaryotic organisms.

The number of additional isoforms varies considerably by species. In the human, mouse, and rat proteomes of the total number of entries, 26 %, 40 % and 72 % are canonical, respectively. The choice of including isoforms is related to the search algorithm and experimental goals. For instance, if differentiating isoforms is relevant, they should be included otherwise they will not be detected. In cases where isoforms are present in the FASTA (evident by shared protein names) but these cannot be removed prior to downloading (e.g., California sea lion, *Zalophus californianus*, proteome

UP000515165, release 2022_01), non-redundant FASTA files can be manually generated (i.e., "remove_duplicates.py" via [196]). If possible, retrieving canonical protein sequences via proteomes is the most straight forward approach and in general appropriate for most search algorithms, versus the method of searching and downloading Swiss-Prot and/or TrEMBL entries.

Though FASTA files are the typical input of many search algorithms, UniProt also offers an XML and GFF format download. In contrast to the flat FASTA file format, the XML format includes sequence information as well as associated information like PTMs, which is used in some search algorithms like MetaMorpheus [PMID:26418581?].

Once a protein sequence collection has been selected and retrieved, there is the evergreen question of how to name and report this to others in a way that allows them to reproduce the retrieval. The minimum reporting information is the taxon identified and number of sequences used [193, PMID:23500130?]. The following naming format (and those below) augments this and is suggested for UniProtKB FASTA files (the use of underscores or hyphens is not critical): [common or scientific name]-[taxon id]-uniprot-[swiss-prot/trembl/proteome]-[UP# if used]-[canonical/canonical plus isoform]-[release] example of a Homo sapiens (human) protein fasta from UniProtKB:

Human-9606-uniprot-proteome-UP000005640-canonical-2022_01.fasta

The importance of the taxon identifier has already been described above and is a consistent identifier across time and shared across resources. The choices of Swiss-Prot and TrEMBL in some combination was discussed above, and Proteome can be "proteome", "reference proteome" or "pan-proteome". The proteome identifier ('UP' followed by 9 digits) is conserved across releases, and release information should also be included. A confusing issue to newcomers is what the term "release" means. This is a year_month format (e.g., 2022_01), but it is not the date a FASTA file was downloaded or created, nor does it imply there are monthly updates. This release "date" is a traceable release identifier that is listed on UniProt's website. Including all this information ensures that the exact provenance of a FASTA file is known and allows the FASTA file to be regenerated.

RefSeq

NCBI is a clearing house of numerous types of data and databases. Specific to protein sequence collections, NCBI Reference Sequence Database (RefSeq) provides annotated genomes across the Tree of Life. The newly developed NCBI Datasets portal [197] is the preferred method for accessing the myriad of NCBI data products, though protein sequence collections can also be retrieved from RefSeq directly[198,199]. Like UniProt described above, most of the additional functionality and information available through NCBI Datasets and RefSeq will not be described here, although the Eukaryotic RefSeq annotation dashboard [200] is a noteworthy resource to monitor the progress of new or re-annotations. We recommend exploring the resources available from NCBI [201], utilizing their tutorials and help requests.

RefSeq is akin to the "proteome" sequence collection from UniProtKB, where a release is based on a single genome assembly. If a more complete genome assembly is deposited or additional secondary evidence (e.g., RNA sequencing) is deposited, RefSeq can update the annotation with a new annotation release. Every annotation release will have an annotation report that contains information on the underlying genome assembly, the new genome annotation, secondary evidence used, and various statistics about what was updated. The current annotation release is referred to as the "reference annotation", but each annotation is numbered sequentially starting at 100 (the first release). Certain species are on scheduled re-annotation, like human and mouse, while other species are updated as needed based on new data and community feedback (ex. release 100 of taxon 9704 was in 2018, but a more contiguous genome assembly resulted in re-annotation to release 101 in

2020). This general process for new and existing species is described in Heck and Neely [PMID:32786681?].

Since RefSeq is genome assembly-centric, its protein sequence collections are retrieved for each species. This contrasts with being able to use a higher-level taxon identifier like 40674 (Mammalia) in UniProt to retrieve a single FASTA. To accomplish this same search in NCBI Datasets requires a Mammalia search, followed by browsing all 2083 genomes and then filtering the results to reference genomes with annotations, and those resulting 188 could be bulk downloaded, though this will still be 188 individual FASTA files. It is possible to download a single FASTA from an upper-level taxon identifier using the NCBI Taxonomy Browser, though this service may be redundant with the new NCBI Datasets portal. Given the constant development of NCBI Datasets, these functionalities may change, but the general RefSeq philosophy of single species FASTA should be kept in mind. Likewise, when retrieving genome annotations there is no ability to specify canonical entries only, but it is possible to use computational tools to remove redundant entries ("remove_duplicates.py" from [196]).

Similar to the UniProtKB FASTA file naming suggestion, the following naming format is suggested for RefSeq protein sequence collection FASTA (the use of underscores or hyphens is not critical): [common or scientific name]-[taxon id]-refseq-[release number] example of a Equus caballus (horse) protein FASTA from RefSeq: Equus_caballus-9796-refseq-103.fasta The release number starts at 100 and is consecutively numbered. Note, the human releases only recently began following this consecutive numbering for Release 110, and previously had a much longer number to be included (e.g., NCBI Release 109.20211119). Also, in a few species (Human and Chinese hamster, currently), there is a reference and an alternate assembly, both with an available annotation. In these cases, including the underlying assembly identifier would be needed. Note that when you retrieve the protein FASTA from NCBI it will include two more identifiers that aren't required in the file name since it can be determined from the taxon identifier and release number. These are the genome assembly used (this is generated by the depositor and follows no naming scheme) and the RefSeq identifier (GCF followed by a number string). These aren't essential for FASTA naming, but are for comparing between UniProt, RefSeq and Ensembl when the same underlying assembly is used (or not, indicating how up to date one is versus the other).

Ensembl

There are two main web portals for Ensembl sequence collections: the Ensembl genome browser [202] has vertebrate organisms and the Ensemble Genome project [203] has specific web portals for different non-vertebrate branches of the Tree of Life. This contrasts with NCBI and UniProt where all branches are centrally available. Recently, Ensembl has created a new portal "Rapid Release" focusing on quickly making annotations available (replacing the "Pre-Ensemble" portal), albeit without the full functionality of the primary Ensembl resources. Overall, Ensembl provides diverse comparative and genomic tools that should be explored, but, specific to this discussion, they provide species-specific genome annotation products similar to RefSeq.

To retrieve a protein sequence collection from Ensemble at any of the portals, a species can be searched using a name, which will then have taxon identifier displayed (but searching by identifier is not readily apparent). From the results you can select your species and follow links for genome annotation. Caution should be used when browsing the annotation products since the protein coding sequence (abbreviated "cds") annotations are nucleic acid sequences (a useable via 3-frame translation if using certain software), while actual translated peptide sequences are in the "pep" folders. The pep folders contain file names with "ab initio" and "all" in the FASTA file names (file extensions are "fa" for FASTA and "gz" indicating gzip compression algorithm), while there may only be one pep product for certain species in the "Rapid Release" portal. The "ab initio" FASTA files contain mostly predicted gene products. The "all" FASTA files are the usable protein sequence collections. Ensembl FASTA files usually have some protein sequence redundancy.

Ensembl provides a release number for all the databases within each portal. Similar to the UniProt file naming suggestion, the following naming format is suggested for Ensembl protein sequence collection FASTA (the use of underscores or hyphens is not critical):

[common or scientific name]-[taxon id]-ensembl-[abinitio/all]-[rapid]-[release number]

example of a Sus scrofa (pig) protein FASTA from Ensembl:

Pig-9823-ensembl-all-106.fasta

Similar to the FASTA download from RefSeq, the downloaded file name can include additional identifying information related to the underlying genome assembly. Again, this is not required for labeling, but is useful to easily compare assembly versions.

Since much of the data from Ensembl is also regularly processed into UniProt, using UniProt sequence collections instead may be preferred. That said, they are not on the same release schedule nor will the FASTA files contain the same proteins. Ensembl sequences still must go through the established protein sequence pipeline at UniProt to remove redundancy and conform to UniProt accession and FASTA header formats. Moreover, the gene-centric and comparative tools built into Ensembl may be more experimentally appropriate and using an Ensembl protein sequence collection can better leverage those tools.

Other resources

There are other locations of protein sequence collections, and these will likewise have different FASTA file formatting; sequences may have unusual characters, and formats of accessions and FASTA header lines may need to be reformatted to be compatible with search software. These alternatives include institutes like the Joint Genome Institute's microbial genome clearing house, species-specific community resource (e.g., PomBase, FlyBase, WormBase, TryTrypDB, etc.), and one-off websites tenuously hosting in-house annotations. It is preferred to use protein sequence collection from the main three sources described here, since provenance can be tracked, and versions maintained. It is beyond the scope of this discussion to address other genome annotation resources, how they are versioned, or the best way to describe FASTA files retrieved from those sources. In these cases, defaulting to the minimum requirements of listing number of entries and supplying the FASTA along with data are necessary.

Contaminants

Samples are rarely comprised of only proteins from the species of interest. There can be protein contamination during sample collection or processing. This may include proteins from human skin, wool from clothing, particles from latex, or even porcine trypsin itself, all of which contain proteins that can be digested along with the intended sample and analyzed in the mass spectrometer. Avoiding unwanted matching of mass spectra originating from contaminant proteins to the cellular proteins due to sequence similarities is important to the identification and quantitation of as many cellular proteins as possible. To avoid random matching, repositories of supplementary sequences for contaminant proteins have been added to a reference database for MS data searches. Appending a contaminants database to the reference database allows the identification of peptides that are not exclusive to one species. Peptides that are exclusive to the organism of interest are used to calculate abundance to avoid inflated quantitative results due to potential contaminant peptides.

As early as 2004, The Global Proteome Machine was providing a protein sequence collection of these common Repository of Adventitious Proteins (cRAP), while another contaminant list was published in

2008 [[PMID:18790129?](#)]. The current cRAP version (v1.0) was described in 2012 [[204](#)] and is still widely in use today. cRAP is the contaminant protein list used in nearly all modern database searching software, though the documentation, versioning or updating of many of these “built-in” contaminant sequence collections is difficult to follow. There is also another contaminant sequence collection distributed with MaxQuant. Together, the cRAP and MaxQuant contaminant protein sequence collections are found in some form across most software, including MetaMorpheus and Philosopher (available in FragPipe) [[PMID:32669682?](#)]. This list of known frequently contaminating proteins can either be automatically included by the software or can be retrieved as a FASTA to be used along with the primary search FASTA(s). Recently the Hao Lab has revisited these common contaminant sequences in an effort to update the protein sequences, test their utility on experimental data, and add or remove entries [[205](#)].

In addition to these environmentally unintended contaminants, there are known contaminants that also have available protein sequence collections (or can be generated using the steps above) and should be included in the search space. These can include the media cells were grown in (e.g., fetal bovine serum [[PMID:20641139?](#),[PMID:33532042?](#)], food fed to cells/animals (e.g., *Caenorhabditis elegans* grown on *Escherichia coli*) or known non-specific binders in affinity purification (i.e., CRAPome [[PMID:23921808?](#)]). The common Repository of Fetal Bovine Serum Proteins (cRFP)[[PMID:31475827?](#)] are protein lists of common protein contaminants and fetal serum bovine sequences used to reduced the number of falsely identified proteins in cell culture experiments. Cells washed or cultured in contaminant free media before harvest or the collection of secreted proteins depletes most high abundance contaminant proteins but the sequence similarity between contaminant and secreted proteins can cause false identifications and overestimation of the true protein abundance leading to wasted resources and time on validating false leads. As emphasized throughout this section, accurately defining the search space is essential for accurate results and, especially in the case of contaminants, requires knowledge of the experiment and sample processing to adequately define possible background proteins.

Choosing the right database

Proteomics data analysis requires carefully matching the search space (defined by the database choice) with the expected proteins. A properly chosen database will minimize false positives and false negatives. Choosing a database that is too large will increase the number of false positives, or decoy hits, which in turn will reduce the total number of identifiable proteins. For this reason it is ill advised to search against all possible protein sequences ever predicted from any genomic sequence. On the other hand, choosing a database that is too small may increase false negatives, or missed protein identifications, because in order for a protein to be identified it must be present in the database. Thus, proteomics practitioners must do their best to predict the proteins that might be in their sample before they analyze their data.

Proteomics data analysis requires carefully aligning the search space with the expected proteome and the statistical approach of the search algorithm. Search algorithms can self-correct when a database is overly large such that higher identity thresholds are required for identification to minimize false positives (e.g., Mascot), while smaller experiment-specific search spaces (also referred to as “subsets”) can have unintended effects on false positives if not managed appropriately [[PMID:34236864?](#),[PMID:30560673?](#),[PMID:26125591?](#)] or may even improve protein identifications [[PMID:27975281?](#)]. Whether to employ a search space that is sample-specific (i.e., subset), species-specific (with only canonical proteins, described below), exhaustive species-specific (including all isoforms), or even larger clade-level protein sequence set (e.g., the over 14 million protein sequences associated with Fungi, taxon identifier 4751) is a complex issue that is experiment and software dependent. Moreover, in cases where no species-specific protein sequence collection exists, homology-based searching can be used (as described in [[PMID:32786681?](#)]). In each of these cases, proteomics practitioners must understand their specific experimental sample and search algorithm in

order to know how to best define the search space, which is essential to yielding accurate results. See more discussion of database choice in the following section.

Biological Interpretation

The most common untargeted proteomics experiment will produce a list of proteins or peptides of interest which require further validation and biological interpretation. This list usually results from statistical data analysis; the typical output of differentially expressed proteins usually contains hundreds of hits. In this section, we aim to present a concise overview of how proteomic data can be effectively contextualised and used to generate new hypotheses.

The simplest approach is to start manual lookup of every protein in the list to uncover groups that function together. Starting with a list of hundreds of protein changes, a smaller list can be prioritized by considering the level of significance and effect size. For example, proteins with the smallest p-values (significance) and largest abundance fold-changes (effect size). It is tempting to focus on proteins with the most extreme fold changes. In this case, the assumption is that the more significant the fold change (in either direction, up- or down-regulation), the higher the impact of those proteins on cellular behavior. This assumption is not always valid because protein signal in MS depends on abundance. The manual data interpretation approach is typically infeasible due to the number of proteins that would need to be individually looked up one-by-one.

A better strategy is to use computational methods. These methods may consider the whole list of proteins including some ranking by significance or fold change. One common interpretation method is to construct a protein network, which then lends itself to network analyses. Another method is to consider functional enrichment through annotation databases. These databases offer insights by examining the enrichment of certain functional annotations amongst the interesting proteins. Secondly, one could consider other evolutionary, structurally or regulatory based methods to identify interpretation of the data. To fully interpret analysis, it may be required to perform or examine other data such as data from biophysical, biochemical and alternative proteomic approaches. Finally, the data can further be interpreted using multi-omic, native or clinical approaches. Below we summarize these approaches and point out potential pitfalls with these methods.

Constructing a protein network

A network is a representation of the relations between objects. Nodes are the entities of the network (e.g., users of a social platform, train stations, proteins), while edges are the connections between them (e.g., friendship, routes, and protein interactions, respectively). In the case of protein-protein interactions, the evidence for the functional associations between proteins can be obtained experimentally. For example, co-immunoprecipitation, crosslinking, and proximity labeling can be used to reveal physical interactions [[PMID:34046695?](#)]. The data is presented in a table with nodes and edges (e.g., “protein A interacts with protein B”) from which the network can be constructed. A considerable wealth of protein-protein association data is stored in free databases like IntAct, which contain interactions derived from literature curation or direct user submissions [[PMID:34761267?](#)]. Protein interactions can also be predicted by classifiers that consider many features, like orthology and co-localization, to produce a posterior odds ratio of interaction [[PMID:21572181?](#), [PMID:33237311?](#)]. Finally, large repositories like STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) collect and integrate protein-protein interaction data from several databases [[PMID:33237311?](#)]. STRING also provides a web-based interface to survey the data, and users only have to feed a search box with the identifiers of the protein(s) of interest. STRING will retrieve the network and show the evidence supporting each interaction. Importantly, these databases do not indicate the direction of the interaction, so they produce undirected networks. If edges have directions (e.g., A influences B and not vice versa), then the network is directed. Signaling pathways are examples of directed graphs.

Network analysis

Network analysis is a group of techniques that explore and investigate the network, yielding valuable knowledge about its structure and unveiling key players regulating the flow of information. One of the first steps in network analysis relates to centrality measurements. Centralities are indicators of the relative importance of a node corresponding to its position in the network, and each centrality measure provides new insights to interpret the data in new ways [[PMID:27616995?](#),[PMID:30064421?](#)].

Degree centrality

The degree of a node measures the number of edges incident to that node. Nodes with a high degree interact with many other nodes, called first neighbors. In particular, the node degree distribution in protein networks is highly skewed, with most nodes having a low degree and a few having high degrees, known as hubs. Hubs are usually regulatory proteins, being notable examples oncogenes and transcription factors. Moreover, hubs are attractive targets for directed interventions, as their alteration has a profound effect on the stability of the network [[PMID:16751849?](#)].

Closeness centrality

The route from one node to another is a path, and the shortest path is the one connecting them in the least amount of steps. Closeness centrality is the inverse of the average length of a node's shortest paths to all other nodes in the network. Nodes with a high closeness score have the shortest distances to all the others, so closeness centrality calculations detect nodes that can spread information very efficiently, as they are in a better position in the network for this task [[206/](#),[PMID:15616139?](#)].

Betweenness centrality

This centrality index is related to the amount of shortest paths transversing a node. Nodes with a high betweenness centrality usually bridge different parts of the network and strongly influence the flow of information, as they lie in communication paths. These connector hubs (or bottlenecks) are also interesting for follow-up experiments because their removal can disconnect different regions of the network [[PMID:17447836?](#)].

Centrality measurements add new layers of information and allow for ranking differentially expressed proteins apart from their fold-change in abundance. Figure 4X depicts a simple network consisting of proteins A to L, with A having the highest fold change (10) and L the lowest (2). In Panel A, the fill color for the nodes indicates this metric, where it can be easily seen that A stands out. However, protein A is a peripheral protein, only interacting with B. In Panel B, nodes are colored according to node degree. Clearly, protein F has the highest number of interactions and is also the closest to all other nodes, which can be appreciated when nodes are colored according to closeness centrality (Panel C). On the other hand, protein G acts as a bridge between two regions of the network and thus, has the highest betweenness centrality (Panel D). Except for fold change, node A has the lowest indices, and it will be up to the researcher to decide whether this protein warrants further examination.

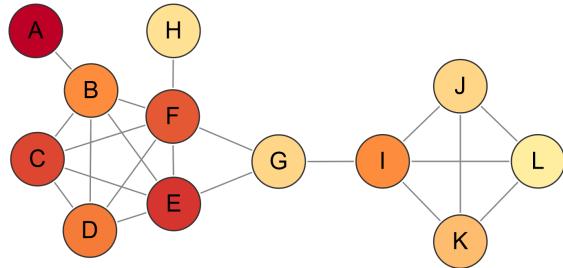
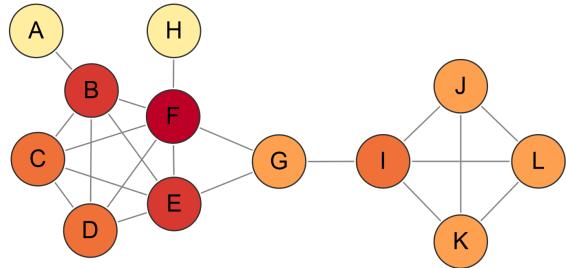
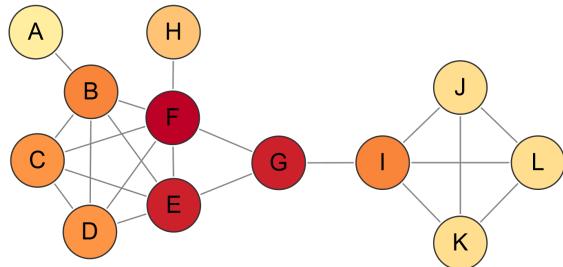
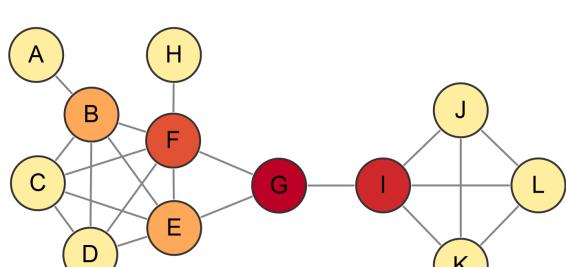
A) Fold change**B) Degree****C) Closeness centrality****D) Betweenness centrality**

Figure 4: Analysis of a simple network using different centrality measurements. Nodes are colored according to each metric using a yellow-to-red gradient (yellow: lowest value, red: highest value). Network visualization and analysis were performed in Cytoscape.

Network clustering

In the small network presented in Figure 4X, two groups of densely connected nodes exist. This topology suggests that these communities (or “clusters”) work together or participate in a protein complex. Dividing a network into clusters helps identify underlying relationships among nodes, which is especially useful in large networks. In a broad sense, network clustering groups nodes according to a topological property, generally interconnectedness. There are many network clustering algorithms, each with its own merits and approaches [PMID:20158874?, PMID:17087821?]. The MCL (Markov Clustering) algorithm is suitable for protein networks in most situations . On the other hand, the Molecular COmplex DEtection (MCODE) algorithm helps detect very densely connected nodes, thus unveiling protein complexes [PMID:12525261?]. In this regard, network clustering is useful for tentatively assigning the function of an uncharacterized protein. If the protein appears in a cluster, its function should be closely related to the cluster members, a principle known as “guilty by association.” [PMID:26215734?]

Network visualization

A critical step in network analysis is to display the data in a structured and uncluttered graph. Networks can rapidly become a hairball unamenable to interpretation. Software platforms like Cytoscape can be used to visualize networks orderly by applying layout algorithms and format styles [PMID:14597658?]. Since many of these platforms are open source, community-designed plugins enhance their capabilities. In Cytoscape, the stringApp adds a search bar to query the STRING database with accession numbers or protein names [PMID:36512705?]. The network is directly retrieved into Cytoscape, where its built-in network analyzer can be used to calculate centralities. Moreover, user-defined information, like fold-change values, can be integrated and mapped into the network.

Functional term enrichment analysis: KEGG, String, GO, GSEA, ORA, Reactome, and others

Term enrichment analysis is performed to assess whether particular ‘functional terms’ are over-represented in a list of proteins (e.g. from a proteomics experiment) [207,208,209]. For example, after a differential abundance analysis, we may wish to examine whether there is any shared function amongst the proteins which were determined to have significant changes. The simplest analysis to test whether this subset contains more of any particular functional terms than we would expect given the background of proteins. For example, the Gene Ontology is split into the classes: Cellular Component, Molecular Function and Biological Function and we might be interested as to whether our proteins may be more likely to localize to a particular subcellular niche [PMID:10802651?]. The Cellular Component terms could give us a starting point if this might be the case, by examining if Cellular Component annotations are enriched.

There are a number of databases and tools to perform such analysis, which can even be extended to examine whole pathways, networks, post-translational modification and literature representation. For example, databases such as KEGG [PMID:10592173?], String [PMID:33237311?], Reactome [PMID:31691815] and PhosphoSitePlus [PMID:30445427?] can be used to test or annotate a list of proteins. For example, proteomics analysis of human cardiac 3D microtissue exposed to anthracyclines (drugs used in cancer chemotherapy) unearthed several proteins with altered levels [PMID:33060801?]. Many of these were specifically grouped under GO terms related to mitochondrial dysfunction, indicating the detrimental effects of these drugs on the organelle. GO terms [PMID:10802651?] or descriptors from other annotation libraries (like KEGG [PMID:10592173?] or REACTOME [PMID:34788843?]) can be retrieved from STRING when constructing a network or from other freely available compendiums. We refer to a number of articles on the topics, including tools, reviews and best-practice [210]. The main points from such analysis is that we can obtain an insight about protein function by looking at whether our list of proteins have similar or the same annotations. A number of limitations should be taken into account for interpretation. The first is that proteins that are more abundant are more likely to be studied, measured and examined in the literature. Hence, abundant proteins will have more annotations than less abundant ones. One key part of the analysis is also to correctly select the background set; that is, the universe of protein which our list is being compared against. By including contaminants or proteins that are not expressed in our system within the list, the results may be unfaithful.

We may also have access to our own curated set of annotations derived either computational or experimental. One may be interested in seeing whether we have enrichment of these annotations amongst the differentially abundant proteins. Our list of proteins could be divided into two groups: differentially abundant or not. These groups could be divided into whether they have a particular annotation: yes or no. This information can be summarised in a two-by-two table, to which we can apply a statistical test to examine whether that annotation is enriched within our differentially abundant proteins. One test that could be used is the hypergeometric test, and another would be a Fisher Exact test.

There are many methods for performing functional enrichment analysis on the data, but they can mainly be classified into three categories (Figure XI{fig:enrich_methods}), as follows.

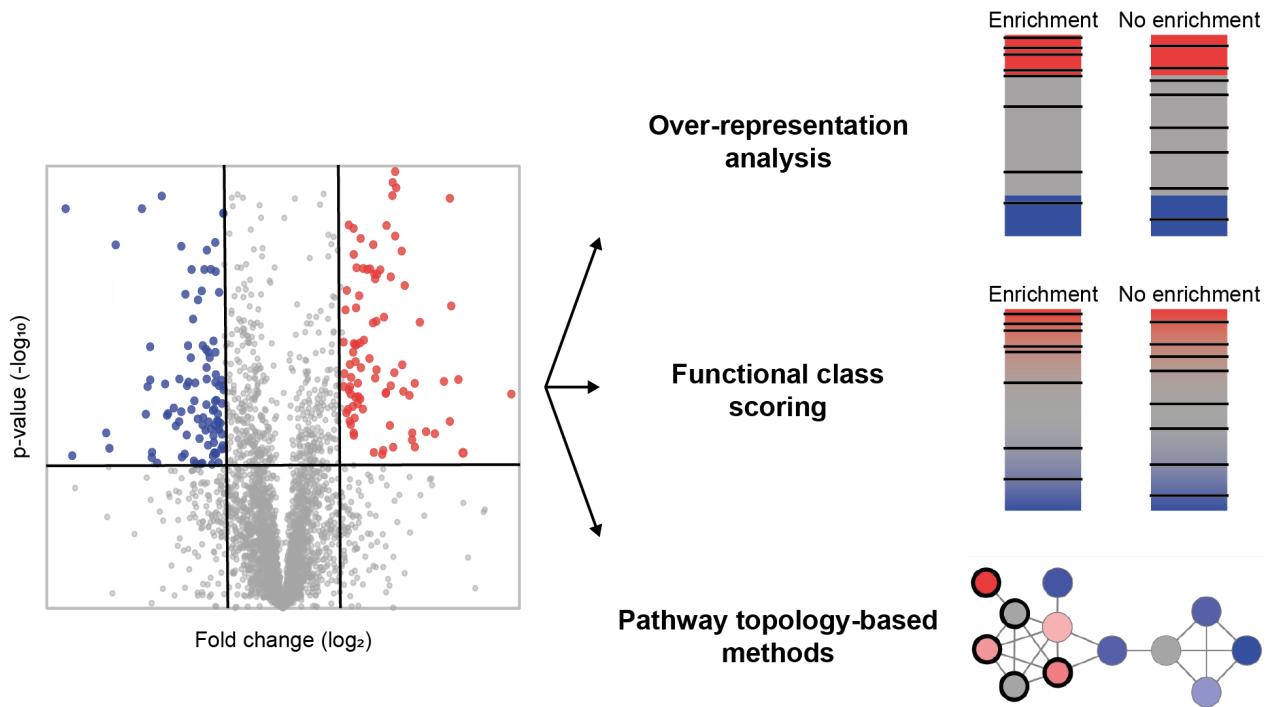


Figure 4: Types of functional enrichment methods. In the volcano plot (left), proteins with altered values are colored blue or red according to arbitrarily chosen cut-off values for significance and fold change. Black bars or thick-bordered nodes indicate members of a GO category.

Over-representation analysis

In a modern proteomics analysis, usually thousands of proteins are identified and quantified. Fold-change and significance thresholds are chosen (e.g., fold-change ≥ 2 and $p \leq 0.05$) to obtain a list of proteins with altered levels among the tested conditions. In over-representation methods, a contingency table is created for every protein set to establish whether proteins with altered abundance show an enrichment or a depletion of the ontology term compared to the background observed proteome [PMID:32695141?]. For example, suppose that 2000 proteins were quantified in a proteomics analysis, being 40 of these members of the set “tricarboxylic acid cycle (TCA).” Also, let us assume that 200 proteins showed altered abundance, with 15 belonging to the TCA set. Then, the contingency table can be constructed as follows:

	Proteins with altered abundance	Proteins with unaltered abundance	Total
Proteins in TCA set	15	25	40
Proteins not in TCA set	185	1775	1960
Total	200	1800	2000

Then, a suitable statistical test is conducted to ascertain if proteins with altered levels are enriched in members of the TCA cycle (in this case, they are; $p < 0.00001$). This is commonly achieved using Fisher’s exact test [211]. The process is then repeated for every set as desired. Since multiple comparisons are tested, p values must be adjusted by a false discovery rate [PMID:35263338?].

Functional class scoring

The caveat of over-representation methods is that they rely on a list of differentially expressed genes or proteins with altered abundance, selected due to arbitrarily chosen cut-off values. For example, if we set a fold change cutoff of 2, a protein with a fold-change of 1.99 would not be included in the analysis. Moreover, several proteins belonging to the same set may have altered levels but are below

the fold change threshold. However, moderate alterations of their abundance as a group could drive the observed phenotype, even more so than a single protein over the cut-off. Functional class scoring strategies aim at countering these limitations by disregarding thresholds altogether. GSEA (Gene Set Enrichment Analysis) is a widely used functional class scoring method in which all detected entities are first ranked according to a quantitative measurement (fold change, p-value, or their combination) [[PMID:16199517?](#)]. Then, the distribution of members of a set is obtained. A scoring scheme based on the Kolmogorov – Smirnov test is used to assess whether there is an enrichment of the category towards the top or bottom of the ranked list.

Pathway topology-based methods

Both methods mentioned above do not consider the functional relationships among proteins put forth by network analysis; i.e., they assume functional independence. Topology-based enrichment methods incorporate this information by, for example, assigning an importance value to a set when its members also participate in a pathway or cluster together in a network [[PMID:24133454?](#)]. In Figure {fig:enrich_methods}XI, it can be shown that topology-based methods consider non-significant hits (grey nodes) that other strategies may not pick up, due to their position in a network.

Other computational approaches: Network analysis, Isoform correlation analysis, AlphaFold, BLAST, protein language models

Additional computational analysis of a list of interesting proteins may uncover additional substructure, correlation or biologically useful hypothesis. Building a network between the proteins based on the experiments performed, might be a useful approach to identify additional structure. For example, co-expression network analysis can be used to build a network from these proteins [[212](#)]. In these networks, proteins are nodes and edges describe relationships between those proteins. Network-specific methods can then be applied, such as community detection algorithms which could uncover clusters of proteins with shared functions [[213,214](#)].

One way the proteome generates complexity is through alternative-splicing, which results in protein isoforms [[215](#)].

Recently, a number of tools have been proposed to identify peptide isoforms that are quantitatively different across conditions by using a principle called peptide correlation analysis [[216, PMID:33325715?](#)]. The idea is that the quantitative behavior of peptides should match each other. If there are subgroups that behave coherently within the group but not across groups suggest that peptide may have come from a different proteoform. These approaches can be used to identify specific proteoforms that are functional across different conditions.

For many, a protein's structure reveals important functional details [[PMID:18429251?](#)]. There are a plethora of approaches to predict a protein's structure [[217,218, PMID:34282049?](#)]. Recently, AlphaFold and RosettaFold have become dominant methods for predicting protein structures with high resolution [[218, PMID:34282049?](#)]. If intrinsically disordered domains are of particular interest, methods explicitly designed for this task are recommended [[PMID:23203878?](#)]. Once a structure is obtained more elaborate computational methods might be useful such as docking or molecular dynamics [[219](#)]. These approaches can give insight into how protein or molecules fit together and the dynamics of a protein's structure (conformational heterogeneity). A complete discussion of these topics is beyond the scope of this section.

Another way to obtain insights into a protein function is to look for protein with similar sequences or motifs. Using BLAST, a sequence alignment tool, one can align two or more protein sequences and determine their level of similarity [[PMID:2231712?](#)]. For example, if a human protein of unknown

function has a similar sequence to a yeast protein with known function this may be a starting place for the putative function of that protein.

Novel approaches to representing the similarity of proteins have proved successful at predicting the functional properties of proteins. Protein language models seek to learn “representation” of proteins, these are usually numerical vectors that represent a protein sequence [221]. Abstractly, these vectors preserve protein similarity or a notion of “proteinness”. This usually means that two proteins that have a close vector may share similarities in protein function. These representations are also advantageous because they can easily become the inputs for machine learning algorithms to predict valuable protein properties; for example, thermal stability values [[PMID:32133759?](#)], protein-protein binding affinities [222], secondary protein structure and more [[DOI:10.1073/pnas.201623911?](#)].

Orthogonal experimental methods

The importance of orthogonal experimental validation

The computational workflows to interpret mass spectrometry data are sophisticated, powerful tools, but also show important limitations and caveats due to their dependence on limited prior knowledge, specific experimental parameters or data quality restraints (see section ‘Analysis of Raw Data’). These inherent biases can give rise to ambiguous or spurious interpretation of the data even when these workflows are applied correctly and to the best of the experimenter’s knowledge. Therefore, researchers will oftentimes be asked by scientific journals to provide independent orthogonal validation of their proteomics data and not performing such can be a major roadblock in the publication process.

The aim of validating data obtained by proteomics approaches should always be two-fold by demonstrating that the conclusions arrived at by proteomics data acquisition and analysis are, firstly, valid and, secondly, relevant. Depending on the question at hand, researchers can draw on an overabundance of techniques to validate MS-derived hypotheses in appropriate cellular, organismal or in vitro models. In the the following paragraphs we aim to present only a high-level, stringent, non-exhaustive selection of orthogonal validation approaches and emphasise the importance of implementing assays that challenge assumptions gained from proteomics data analysis pipelines.

Before embarking on orthogonal validation of any hit, the success of the experiment should be established by assessing (internal) positive controls. Internal positive controls can be proteins whose behaviour under the experimental conditions applied can be deduced from prior knowledge (i.e. the scientific literature or public databases). Once the expected changes in internal controls have been confirmed by computational analysis (see the above section), the orthogonal experimental validation of novel, perhaps unexpected findings can begin.

Orthogonal validation of new insights obtained from quantitative proteomics experiments can be a very time-consuming process and often requires familiarity with techniques not directly related to proteomics workflows. Given these challenges, the method(s) of choice warrant(s) careful consideration and is highly context-dependent. Importantly, proteomics experiments in one way or another generally yield comprehensive lists of potentially interesting candidate proteins or pathways, the researcher will have to shortlist candidates to be taken forward to the validation stage of the project. Which candidates should you validate by an orthogonal approach and which ones might not require further validation?

In general, candidates representing abundant proteins that show high sequence coverage and are detected with high confidence might not necessarily need extensive orthogonal validation when compared with proteins of intermediate to low abundance that might be more challenging to

faithfully quantify by proteomics alone (i.e. many membrane proteins or transcription factors). Similarly, since the proteome is rarely comprehensively quantified in any single proteomics experiment, proteins of interest (POIs) that are critical for an observed biological change might not be part of the dataset. In these cases, additional, targeted analyses might help to support or discredit proteomics-based hypotheses.

Validation techniques are as manifold as biological questions and discussions thereof may easily fill multiple textbooks. The following sections are therefore merely meant to paint with a broad brush stroke a picture of useful methodologies with which to validate and follow up MS-data derived observations. As this is meant to orient the reader, wherever possible, we will explicitly point out useful literature reviews for a deeper dive into each of these techniques.

General considerations

Once POIs have been selected based on prior agreed-upon selection criteria (i.e. (adjusted) p value and/or fold change thresholds), orthogonal validation experiments should ideally be conducted under physiologically relevant conditions to mitigate artificial and misleading outcomes. Therefore, *in vitro* experiments, while useful to isolate and dissect particular aspects of a biological system, can give highly artificial results as conditions are far removed from the POI's native environment. To investigate the biological function of a protein or pathway, direct genetic manipulation of the biological system at hand (e.g. modulating the expression of a POI by overexpression or knockout-/down experiments) can be minimally invasive when performed correctly. Should the POI be encoded by an essential gene, by definition, a complete and stable knockout might not be advisable [223,224]. In these extreme cases, attenuated expression (i.e. using RNA interference (RNAi) or controlled degradation, see below) rather than complete repression of a gene can be used to probe for protein function. Epitope tagging and/or exogenous expression of a gene of interest can be a powerful approach in assessing PPIs and investigating proteins of low abundance. However, overexpression artifacts are common [DOI? 10.1016/j.mrrev.2017.05.002].

It is not always possible to fully avoid the pleiotropic effects of protein (over-)expression or depletion, but a number of mitigation strategies (i.e. inducible expression, the use of multiple independent RNAi strategies) will be discussed below.

Extensive biochemical characterization of any overexpressed gene is critical to ensure it closely reflects the functions of its endogenous counterpart. These assays might involve assessing protein localization (i.e. by imaging techniques such as microscopy and flow cytometry), protein abundance (i.e. by mass spectrometry or immunoblot analysis) and phenotypic assays where applicable and practical.

Functional genomics techniques in the validation of MS hits

Typical follow-up experiments to validate mass-spectrometry derived insights often involve the acute depletion or induction of a POI and assessing the impact on specific cellular phenotypes. Here we present a selection of methodologies to effectively modulate gene expression and discuss important considerations when planning functional genomics experiments for target validation.

Gene deletion or knockdown to prevent production of a functional protein is a powerful means to interrogate the role of one or more proteins in the phenotype(s) under investigation. To this end, well-established technologies deserving mention at this point are RNA interference (RNAi) in the form of siRNA/shRNA- or miRNA-mediated gene knockdown and CRISPR/Cas9- or TALEN-mediated gene knockout [225]. Since each one of these technologies comes with their own unique advantages and caveats, the approach taken depends on the biological question at hand.

Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas-based gene deletion technologies allow for the targeting of individual genes with relative ease, high efficiency and specificity [226]. When expressed in mammalian cells, the bacterially-derived Cas9 endonuclease can be guided with the help of a short guide RNA (gRNA) to a genomic location of interest, where it creates a DNA double strand break in a highly controlled manner (for a detailed discussion see [227]). The cell's DNA double-stand break repair machinery then introduces base pair insertions or deletions (indels) via non-homologous-end-joining (NHEJ), thus causing missense, and frameshift mutations (i.e. resulting in premature stop codons), leading to premature termination of gene expression or non-functional, aberrant gene products. Similarly, the concomitant provision of a complementary DNA donor template encoding a desired gene modification (i.e. insertion of a stretch of DNA or base pair modification) will trigger homology-directed repair (HDR), resulting in gene knockin or base editing [227].

Practical considerations of CRISPR/Cas9-mediated gene knock-in and base editing will not be addressed in detail but are expertly discussed in [228], [229], [230], [231].

The relative ease-of-use and high efficiency of the CRISPR/Cas9 gene editing technology has rendered it the method of choice for gene manipulation in many fields of cell biology. However, it should be noted that CRISPR/Cas9-mediated gene deletion is not free from off-target effects ([232] for advice on how to minimise these off-target effects). Moreover, long-term depletion (or upregulation) of a POI itself can in some cases have dramatic systemic consequences and constitute an acute selection pressure leading to compensatory stress-induced adaptation that might obfuscate primary loss-of-function phenotypes and pose a substantial hurdle to the interpretability of biological data. As these compensatory mechanisms often manifest with time, controlled, transient genetic manipulation (gene depletion or transgene expression) is advised. Small interfering RNA (siRNA)-mediated knockdown by transient transfection is typically achieved at shorter time frames (24 – 96h), depending on the turnover of the POI. On an even shorter time-scale, targeted, degron-based degradation systems enable depletion of a POI within minutes and further reduce off-target effects, but require the exogenous expression of a transgene and therefore some genetic manipulation. A more comprehensive discussion of a selection of these systems (anchor-away, deGradFP, auxin-inducible degron (AID), dTAG technologies) and their advantages and potential pitfalls is presented in [233].

Multiple eukaryotic and prokaryotic transcription-based systems have been developed that allow for the controlled biosynthesis or depletion of one or more POIs. Amongst these, a popular and dependable choice for mammalian cells are tetracycline-controlled operon systems, which allow up- or downregulation of a POI in the presence of the antibiotic tetracycline or its derivative doxycycline. These systems rely on the insertion of a bacteria-derived Tet operon (TetO) between the promoter and coding sequence of a GOI. In this configuration, the TetO binds a co-expressed Tet-repressor protein blocking transcription of the GOI. When tetracycline is added to the cells, the repressor then dissociates from the operon, thus de-repressing the GOI. Different variations of this potent system exist, allowing for more flexibility in experimental design. For instance, in the Tet-OFF system, the Tet repressor is fused to a eukaryotic transactivator (the chimeric fusion construct is termed tTA) and addition of tetracycline, or the related doxycycline, abolishes TetO binding and thus suppresses transcriptional activation [234]. Alternatively, a mutant form of tTA (rtTA) binds the TetO only in the presence of tetracycline, allowing for tetracycline-induced gene expression. For a detailed discussion of these systems, we refer the reader to an excellent review [235].

When generating stable expression cell lines, being able to precisely control the genomic integration site of the transgene reduces overall genetic heterogeneity in a cell population and thereby reduces potential off-target or pleiotropic effects. This ability is realised in the FlpIn-T-REx technology which harnesses Flp-recombinase mediated DNA recombination at a strictly defined genomic locus (the FRT site) [236]. Site-directed isogenic integration of any GOI at the FRT site, which is under a tetracycline-inducible promoter and a hygromycin resistance gene, allows for facile generation of

tetracycline/doxycycline-inducible isogenic expression cell lines with minimal leaky expression (for an example, see [237]).

Validation and interpretation of protein abundance changes

To validate protein abundance changes observed by quantitative bottom-up proteomics or simply assess the success of targeted genetic manipulation as part of an orthogonal follow-up experiment (see above), the experimenter typically resorts to antibody-based techniques such as immunoblotting analysis or immunofluorescence and immunohistological imaging of POIs. The latter also allows for validation of protein expression and localisation in intact tissue or cells. However, these semi-quantitative methods are strongly influenced by the quality of the antibodies used and might not be sensitive enough to detect small changes in protein levels. In this case, more accurate orthogonal quantitation of proteins might be achieved by stable isotope labelling (SILAC/TMT/iTRAQ) and/or SRM/PRM (see section ‘Experiment Types’). SDS-PAGE and immunoblot analysis are powerful and facile low-throughput tools to quickly validate protein abundance changes. However, short of introducing epitope tags to the endogenous POI, the success of immunoblotting is contingent on the availability of specific antibodies, which can present a formidable problem when investigating poorly characterised proteins or working with model organisms for which the commercial availability of specific antibodies is limited (this is particularly problematic for ‘unconventional’ or even well-established model organisms such as yeast). A detailed discussion of the strengths and pitfalls of immunoblotting for validation of semi-quantitative proteomics data can be found in an excellent review by Handler **et al.** [238].

Protein abundance changes detected in a proteomics experiment can be the result of a range of different cellular processes. The abundance of a protein in a complex sample (e.g. cell lysate or biological fluid) directly reflects a combination of the protein’s intrinsic stability and the translational rate under the conditions of interest.

Both protein stability as well as gene expression activity can be quantified independently. Altered protein stability might be a direct consequence of specific or global changes in protein turnover. Radioisotope labelling is a well-established, accurate way to monitor protein synthesis, maturation and turnover [239], [240]. This ‘pulse-chase’ methodology relies on the incorporation (‘pulsing’) of radioisotopes (typically ^{35}S -labelled cysteine and methionine) into de-novo synthesised proteins. Upon withdrawal of the labelled amino acids from the culture medium, the decay of signal is monitored over time (‘the chase’) by SDS-PAGE and phosphoimaging, resulting in a temporal readout of protein abundances. The advantage of this technology is that a subpopulation (newly synthesised proteins) can be monitored directly, giving an accurate assessment of protein stability. Once a change in protein stability has been validated, the underlying mechanisms can be addressed by inhibiting protein degradation pathways; prominently proteasome-mediated degradation (using specific proteasome inhibitors such as bortezomib/velcade or MG132), autophagy (pharmacologically inhibiting autophagic flux) or degradation by proteases (using protease inhibitors). The type of radiolabeling described above is relatively labor-intensive, of low-throughput and has the obvious disadvantage of requiring radioactive material, which needs to be handled under strict safety precautions. Moreover, it critically depends on the presence of one or more methionines and/or cysteines in the POIs.

It is also possible to measure protein stability within complex protein mixtures (i.e. cell lysates or biological fluids) using an array of specialized mass spectrometry techniques as discussed in [241] and [69].

For purified proteins, well-established *in vitro* spectrometric and calorimetric methods such as circular dichroism, differential scanning calorimetry or differential scanning fluorometry can be used, but the relatively high sample amounts might be restrictive.

Finally, gene expression changes can also be determined with high fidelity using quantitative real-time PCR (qRT-PCR) or RNA-Seq can measure changes in gene transcription or mRNA turnover (for an extensive discussion of both technologies, please see [242] and [243], respectively).

Validation of protein-protein interactions

The interaction of a protein with other proteins determines its function. Protein-protein interactions (PPIs) can be either mostly static (i.e. core subunits of a protein complex) or dynamic, varying with cellular state (i.e. cell cycle phase or cellular stress responses, posttranslational modifications) or environmental factors (i.e. availability of nutrients, presence of extracellular ligands of cell-surface receptors). Therefore, any given protein can typically bind a range of interaction partners in a spatially and temporally restricted manner, thus forming complex PPI networks (the interactome of a protein). The method of choice to experimentally examine altered PPI states depends on the model system and biological question (i.e. purified proteins vs complex protein mixtures, monitoring of PPIs in live cells or cell lysate etc). Popular methods for the validation of PPIs *in vivo* include protein fragment complementation (split protein systems), 2-hybrid assays (mammalian, yeast and bacterial), proximity ligation, proximity labelling and FRET / BRET. Protein fragment complementation assays rely on the principle that the two self-associating halves of reporter proteins can be expressed in an inactive form but when in spatial proximity bind one another to complement the functional, active reporter. When these split reporters are fused to two interacting proteins (so-called bait and prey proteins), the binding of bait to prey induces the spatial restriction needed to fully complement the reporter. Commonly used reporter complementation systems are split fluorescent proteins (i.e. GFP, YFP) [244], ubiquitin [245], luciferase [246], TEV protease [247], beta-lactamase [248], beta-galactosidase, Gal4, or DHFR [249]. The resulting functional readout of these complementation system depends on which split reporter is used. In general, the split luciferase system shows enhanced sensitivity over fluorescence-based systems as background luminescence is low.

Two-hybrid assays are based on a similar functional complementation strategy as fragment complementation systems. Conventionally, two self-complementing transcription factor fragments are fused to bait and prey proteins, respectively, leading to the restoration of a functional transcription factor only upon prey-bait interaction. The complemented transcription factor then induces the expression of a reporter gene that can be measured. Multiple variations of this system abound for different model organisms, but they almost always involve transcriptional activation or repression of a reporter gene ([250] for a detailed discussion).

The yeast-2-hybrid system (Y2H) is deserving of mention here as it had been the very first 2-hybrid system established [251] and has ever since proven to be extremely versatile (multiple auxotrophic reporters and markers of phenotypic sensitivity available), cheap, lends itself to functional high-throughput screening and variants have been developed that allow for the investigation of membrane-protein interactions (i.e. membrane Y2H) [252], [250].

Despite the many advantages the Y2H offers, critical drawbacks include the potential of misfolding of bait and prey proteins when fused to a complementation reporter, expression at non-physiological levels, the lack of control over posttranslational modifications that might be important for the PPI under investigation, and the potential requirement of kingdom- or species-specific folding factors for the bait/prey under investigation (i.e. when probing PPI of mammalian proteins in Y2H). Principles of the Y2H technology have also been adapted to mammalian systems, which circumvent some of the aforementioned drawbacks of Y2H [253].

Perhaps the most commonly applied method of detecting and validating PPIs *in vitro* is affinity purification (AP, also known as affinity chromatography) of co-immunoprecipitation (Co-IP) either coupled with SDS-PAGE/immunoblotting or mass spectrometry to determine the identity of interacting proteins. AP typically relies on the isolation of a transgenic POI by an epitope tag (using

epitope-specific matrix-conjugated proteins (antibodies or epitope-binding proteins)), while Co-IP harnesses specific antibodies directly targeting the POI. Specific interactors are expected to be enriched compared to the negative control (i.e an isotype control antibody, a knockout cell line or empty matrix). AP is not solely restricted to detecting PPIs, but can also be adapted to protein interactions with other biomolecules such as RNA [254]. It should be noted that AP and Co-IP can return multiple potential binding partners, many of which might be artefactual due to loss of cellular compartmentalisation during sample preparation.

To reduce the probability of such artefacts and increase the confidence of a specific interaction, reciprocal affinity purification (by pulldown of each interaction partner) or *in situ* imaging might be performed (i.e. using fluorescence resonance energy transfer (FRET) [255], split-protein systems [256], proximity ligation assay [257] and immunofluorescence microscopy).

Forster and bioluminescence resonance energy transfer (FRET / BRET) can be used for *in situ* visualisation of protein proximities and therefore PPIs. In FRET, non-radiative energy transfer between donor and receptor chromophores (each fused to prey and bait proteins, respectively), results in the emission of a characteristic fluorescence signal only when both prey and bait are in very close proximity (1-10 nm distance) and a suitable light source for donor excitation is provided [258].

The underlying principle of BRET is similar to that of FRET but with the exception of using a chemical substrate which activates bioluminescent donor, such as luciferase, resulting in energy transfer to a fluorescent acceptor molecule [259], [260]. The main advantages of BRET over FRET are independence from an external light source (which can result in photobleaching), but requires at least one of the POIs to be fused to the donor (while in FRET, donor and acceptor can be chemically conjugated to POI-specific antibodies) [259]. FRET can be particularly useful in investigating cell surface protein interactions when using specific antibodies conjugated to donor and acceptor probes as antibodies are not cell-permeable and therefore restricted to targets presented on the cell surface in the absence of membrane permeabilization agents. Other fluorescence-based PPI assays encompass Fluorescence correlation spectroscopy (FCS) and fluorescence cross-correlation spectroscopy (FCCS). These methods use small volumes of fluorescently labelled proteins and can determine their diffusion coefficients, which change in when proteins form a complex [261].

Proximity labelling methods (Proximity ligation and enzymatic proximity labelling (BirA, APEX2, HRP) can surveil labile or transient interaction in live cells in a high-throughput format when coupled with target identification by MS [262], [263]. These approaches harness a biotin ligase (i.e. BirA, Biold2, AirID, BASU, APEX2, HRP) fused to a POI whose interactome is to be determined. In the presence of biotin (for BirA, Biold2, AirID, BASU, APEX2 and HRP) or a biotin-phenol derivative (for APEX2), the biotin ligase will activate the biotin(-phenol) which then covalently biotinylates any protein in close proximity. The activated biotin has a short half-life, ensuring that the effective labelling radius is typically restricted to approximately 10 nm. Biotinylated proteins are isolated by affinity purification with streptavidin-conjugated beads and identified by mass spectrometry or SDS-PAGE/immunoblotting. TurboID, miniTurboID and ultraID, promiscuous biotin ligases faster than BirA, have been developed allowing for shorter treatment times and decreased background signal. The choice of a biotin ligase variant depends on the POI and experimental setup, but in general HRP does not work in cytoplasmic environments where conditions are chemically reducing, but is suitable for labelling proteins extracellular face of the plasma membrane or in the endoplasmic reticulum and golgi apparatus. While TurboID and similar variants have fast kinetics, they can cause depletion of endogenous biotin and therefore cytotoxicity. A major drawback shared by all variants described above is that they necessitate fusion to the POI, which might alter its physiological behaviour and give rise to false positives or false negatives. Moreover, detecting a biotin-labelled protein does not unequivocally designate it as an interaction partner as spatial proximity to the POI-biotin ligase fusion protein without direct binding can result in biotinylation. The inclusion of controls, such as expression

of the biotinylating enzyme alone in the cellular compartment of interest, is therefore particularly important for enzymatic proximity labelling methods.

The *in situ* proximity ligation assay (PLA) combines the specificity of antibodies with the signal amplification capacity of a DNA polymerase reaction. Here, two antibodies, each conjugated to a short single-strand DNA (ssDNA) tag and each specific to one of the two proteins whose interaction is under investigation, are added to fixed cells or tissue. Once bound to their respective targets and only when in direct proximity, the addition of two connector oligonucleotides complementary to each tag ssDNA tag and phi29 DNA polymerase, triggers isothermal rolling circle amplification, eventually resulting in the generation of continuous stretches of repetitive DNA. These DNA products can then be visualised by *in situ* hybridisation with fluorescently labelled oligonucleotides (see [264] for a detailed discussion). PLA has the advantage of visualising the two interacting proteins in their native environment when high-resolution microscopy is used as a readout.

Chemical crosslinking (XL) of proteins can determine PPIs with amino-acid level resolution, and can thereby give valuable insights into the orientation of two or more proteins relative to one another [265]. Recent technical advances also enabled the visualisation of protein-RNA interaction [266]. Various XL chemistries are available (amine-reactive, sulphydryl and photoreactive crosslinkers; reversible vs irreversible) and cross-linked proteins detected by mass spectrometry [267]. In general, applying XL-MS to a mixture of interacting, purified proteins is preferable to *in situ* XL of complex protein mixtures (i.e. cell lysate) as detection and deconvolution of XL peptides is technically and computationally challenging.

Surface plasmon resonance can accurately measure several key kinetics of PPIs with high accuracy (e.g. association and dissociation kinetics, stoichiometry, affinity) [268]. It relies on the quantification of refractive index changes of polarised light shone onto a sensor chip containing a prey protein immobilised on a metal surface (typically gold). When prey and bait proteins interact, the mass concentration at the metal interface changes, altering the refractive index and SPR angle (intensity of the refracted light).

Methods for protein or peptide fractionation

Protein fractionation * SDS-PAGE

Peptide fractionation * bRP

Tandem Mass Spectrometry and Peptide Fragmentation

Tandem Mass Spectrometry

Why do we need tandem mass spectra data for proteomics? As mass accuracy and resolution increase, the number of potential peptide matches decrease. Still, there are many potential matches to a human peptide with a 5 ppm mass tolerance window (citation). Additional information is required to uniquely assign a signal to a peptide. Early approaches described use of accurate mass and time data to uniquely identify peptides [269]. Most current research, however, has pursued using the pattern of peptide fragment masses helps narrow down the number of potential peptides.

Peptide Fragmentation

how it works Modern proteomics utilizes hybrid mass spectrometers, where hybrid means there are multiple mass analyzers working in harmony. Most hybrid mass spectrometers combine a quadropole for mass selection before a high resolution analyzer, usually a TOF or an orbitrap. In these hybrid mass spectrometers, mass selection happens using the first quadrupole to isolate a small mass range around that mass (usually less than 1 thompson). The selected ions then pass to the collision cell where fragmentation happens.

1. Nomenclature of backbone fragments (abc, xyz, internal, losses).
2. Mechanisms of peptide fragmentation, collision based (CID and HCD, PQD?) versus electron based (ECD and ETD).

Basic Spectral Interpretation

Segway into the section on raw data analysis.

Experiment Design

This section should discuss trade offs and balancing them to design an experiment. 1. constraints: Each experiment will have different constraints, which may include the number of samples needed for analysis, or desire to quantify a specific subset of proteins within a sample. 2. sample size 3. statistics 4. costs

Acknowledgements

The authors thank Phil Wilmarth for helpful input. Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose. This manuscript was written collaboratively using manubot [270].

References

1. **Electrospray Ionization for Mass Spectrometry of Large Biomolecules**
John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, Craig M Whitehouse
Science (1989-10-06) <https://doi.org/cq2q43>
DOI: [10.1126/science.2675315](https://doi.org/science.2675315)
2. **Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry**
Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, T Matsuo
Rapid Communications in Mass Spectrometry (1988-08) <https://doi.org/ffbwrr>
DOI: [10.1002/rcm.1290020802](https://doi.org/rcm.1290020802)
3. **An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics**
Dirk A Wolters, Michael P Washburn, John R Yates
Analytical Chemistry (2001-10-25) <https://doi.org/bn4kq6>
DOI: [10.1021/ac010617e](https://doi.org/ac010617e)
4. **A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry**
Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, Ruedi Aebersold
Analytical Chemistry (2003-07-15) <https://doi.org/b2xv45>
DOI: [10.1021/ac0341261](https://doi.org/ac0341261)
5. **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**
Joshua E Elias, Steven P Gygi
Nature Methods (2007-03) <https://doi.org/djz7fz>
DOI: <https://doi.org/10.1038/nmeth1019>
6. **Mass-spectrometric exploration of proteome structure and function**
Ruedi Aebersold, Matthias Mann
Nature (2016-09) <https://doi.org/f83zqm>
DOI: [10.1038/nature19949](https://doi.org/nature19949)
7. **High-throughput quantitative top-down proteomics**
Kellye A Cupp-Sutton, Si Wu
Molecular Omics (2020) <https://doi.org/gnx98p>
DOI: [10.1039/c9mo00154a](https://doi.org/c9mo00154a) · PMID: [31932818](#) · PMCID: [PMC7529119](#)
8. **Proteoforms as the next proteomics currency**
Lloyd M Smith, Neil L Kelleher
Science (2018-03-09) <https://doi.org/gn6p4x>
DOI: [10.1126/science.aat1884](https://doi.org/science.aat1884) · PMID: [29590032](#) · PMCID: [PMC5944612](#)
9. **The quantitative proteome of a human cell line**
Martin Beck, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, Ruedi Aebersold
Molecular Systems Biology (2011-01) <https://doi.org/drjmfj>
DOI: [10.1038/msb.2011.82](https://doi.org/msb.2011.82) · PMID: [22068332](#) · PMCID: [PMC3261713](#)
10. **A Quantitative Proteome Map of the Human Body**

Lihua Jiang, Meng Wang, Shin Lin, Ruiqi Jian, Xiao Li, Joanne Chan, Guanlan Dong, Huaying Fang, Aaron E Robinson, Michael P Snyder, ... Simona Volpi

Cell (2020-10) <https://doi.org/ghbjrk>

DOI: [10.1016/j.cell.2020.08.036](https://doi.org/10.1016/j.cell.2020.08.036) · PMID: [32916130](https://pubmed.ncbi.nlm.nih.gov/32916130/) · PMCID: [PMC7575058](https://pubmed.ncbi.nlm.nih.gov/PMC7575058/)

11. **Proteoform: a single term describing protein complexity**

Lloyd M Smith, Neil L Kelleher

Nature Methods (2013-02-27) <https://doi.org/gwcc>

DOI: [10.1038/nmeth.2369](https://doi.org/10.1038/nmeth.2369) · PMID: [23443629](https://pubmed.ncbi.nlm.nih.gov/23443629/) · PMCID: [PMC4114032](https://pubmed.ncbi.nlm.nih.gov/PMC4114032/)

12. **Post-translational modifications in proteins: resources, tools and prediction methods**

Shahin Ramazi, Javad Zahiri

Database (2021-01-01) <https://doi.org/gpjz7g>

DOI: [10.1093/database/baab012](https://doi.org/10.1093/database/baab012) · PMID: [33826699](https://pubmed.ncbi.nlm.nih.gov/33826699/) · PMCID: [PMC8040245](https://pubmed.ncbi.nlm.nih.gov/PMC8040245/)

13. **Deciphering protein post-translational modifications using chemical biology tools**

Anne C Conibear

Nature Reviews Chemistry (2020-10-06) <https://doi.org/gqnxhb>

DOI: [10.1038/s41570-020-00223-8](https://doi.org/10.1038/s41570-020-00223-8)

14. **Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions**

Christopher J Oldfield, AKeith Dunker

Annual Review of Biochemistry (2014-06-02) <https://doi.org/gfw6t4>

DOI: [10.1146/annurev-biochem-072711-164947](https://doi.org/10.1146/annurev-biochem-072711-164947)

15. **A Review: Molecular Chaperone-mediated Folding, Unfolding and Disaggregation of Expressed Recombinant Proteins**

Komal Fatima, Fatima Naqvi, Hooria Younas

Cell Biochemistry and Biophysics (2021-02-25) <https://doi.org/gpd4pb>

DOI: [10.1007/s12013-021-00970-5](https://doi.org/10.1007/s12013-021-00970-5)

16. **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**

Yoav Benjamini, Yosef Hochberg

Journal of the Royal Statistical Society: Series B (Methodological) (1995-01)

<https://doi.org/gfpkdx>

DOI: [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)

17. **Large-scale phosphorylation analysis of mouse liver**

Judit Villén, Sean A Beausoleil, Scott A Gerber, Steven P Gygi

Proceedings of the National Academy of Sciences (2007-01-30) <https://doi.org/c7zbpm>

DOI: [10.1073/pnas.0609836104](https://doi.org/10.1073/pnas.0609836104) · PMID: [17242355](https://pubmed.ncbi.nlm.nih.gov/17242355/) · PMCID: [PMC1785252](https://pubmed.ncbi.nlm.nih.gov/PMC1785252/)

18. **PhosphoPep—a phosphoproteome resource for systems biology research in Drosophila Kc167 cells**

Bernd Bodenmiller, Johan Malmstrom, Bertran Gerrits, David Campbell, Henry Lam, Alexander Schmidt, Oliver Rinner, Lukas N Mueller, Paul T Shannon, Patrick G Pedrioli, ... Ruedi Aebersold

Molecular Systems Biology (2007-01) <https://doi.org/fg6rnn>

DOI: [10.1038/msb4100182](https://doi.org/10.1038/msb4100182) · PMID: [17940529](https://pubmed.ncbi.nlm.nih.gov/17940529/) · PMCID: [PMC2063582](https://pubmed.ncbi.nlm.nih.gov/PMC2063582/)

19. **Quantitative Analysis of Phosphotyrosine Signaling Networks Triggered by CD3 and CD28 Costimulation in Jurkat Cells**

Ji-Eun Kim, Forest M White

The Journal of Immunology (2006-03-01) <https://doi.org/gr3m64>

DOI: [10.4049/jimmunol.176.5.2833](https://doi.org/10.4049/jimmunol.176.5.2833)

20. **Quantitative Phosphoproteomic Analysis of T Cell Receptor Signaling Reveals System-Wide Modulation of Protein-Protein Interactions**
Viveka Mayya, Deborah H Lundgren, Sun-Il Hwang, Karim Rezaul, Linfeng Wu, Jimmy K Eng, Vladimir Rodionov, David K Han
Science Signaling (2009-08-18) <https://doi.org/d3czmn>
DOI: [10.1126/scisignal.2000007](https://doi.org/10.1126/scisignal.2000007)
21. **Quantitative phosphoproteomic analysis of signaling network dynamics**
Forest M White
Current Opinion in Biotechnology (2008-08) <https://doi.org/bhd9vx>
DOI: [10.1016/j.copbio.2008.06.006](https://doi.org/10.1016/j.copbio.2008.06.006) · PMID: [18619541](#) · PMCID: [PMC2567239](#)
22. **Glycoproteomics**
Ieva Bagdonaitė, Stacy A Malaker, Daniel A Polasky, Nicholas M Riley, Katrine Schjoldager, Sergey Y Vakhrushev, Adnan Halim, Kiyoko F Aoki-Kinoshita, Alexey I Nesvizhskii, Carolyn R Bertozzi, ... Nichollas E Scott
Nature Reviews Methods Primers (2022-06-23) <https://doi.org/gr3m6w>
DOI: [10.1038/s43586-022-00128-4](https://doi.org/10.1038/s43586-022-00128-4)
23. **Biological roles of glycans**
Ajit Varki
Glycobiology (2016-08-24) <https://doi.org/f9v9j5>
DOI: [10.1093/glycob/cww086](https://doi.org/10.1093/glycob/cww086) · PMID: [27558841](#) · PMCID: [PMC5884436](#)
24. **Glycosylation in Cellular Mechanisms of Health and Disease**
Kazuaki Ohtsubo, Jamey D Marth
Cell (2006-09) <https://doi.org/d2qc7z>
DOI: [10.1016/j.cell.2006.08.019](https://doi.org/10.1016/j.cell.2006.08.019)
25. **Neurological Aspects of Human Glycosylation Disorders**
Hudson H Freeze, Erik A Eklund, Bobby G Ng, Marc C Patterson
Annual Review of Neuroscience (2015-07-08) <https://doi.org/f3pb3r>
DOI: [10.1146/annurev-neuro-071714-034019](https://doi.org/10.1146/annurev-neuro-071714-034019) · PMID: [25840006](#) · PMCID: [PMC4809143](#)
26. **Universal sample preparation method for proteome analysis**
Jacek R Wiśniewski, Alexandre Zougman, Nagarjuna Nagaraj, Matthias Mann
Nature Methods (2009-04-19) <https://doi.org/cfbvh>
DOI: [10.1038/nmeth.1322](https://doi.org/10.1038/nmeth.1322)
27. **Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis**
Alexandre Zougman, Peter J Selby, Rosamonde E Banks
PROTEOMICS (2014-03-26) <https://doi.org/f2rdw4>
DOI: [10.1002/pmic.201300553](https://doi.org/10.1002/pmic.201300553)
28. **Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation***
Tanveer S Batth, MaximAX Tollenaere, Patrick Rüther, Alba Gonzalez-Franquesa, Bhargav S Prabhakar, Simon Bekker-Jensen, Atul S Deshmukh, Jesper V Olsen
Molecular & Cellular Proteomics (2019-05) <https://doi.org/gf3vws>
DOI: [10.1074/mcp.tir118.001270](https://doi.org/10.1074/mcp.tir118.001270) · PMID: [30833379](#) · PMCID: [PMC6495262](#)
29. **S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics**
Milkessa HaileMariam, Rodrigo Vargas Eguez, Harinder Singh, Shiferaw Bekele, Gobena Ameni, Rembert Pieper, Yanbao Yu

30. **Multiplexed CuAAC Suzuki-Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling**
Jian Cao, Lisa M Boatner, Heta S Desai, Nikolas R Burton, Ernest Armenta, Neil J Chan, José O Castellón, Keriann M Backus
Analytical Chemistry (2021-01-20) <https://doi.org/gr3m6q>
DOI: [10.1021/acs.analchem.0c04726](https://doi.org/10.1021/acs.analchem.0c04726) · PMID: [33470097](https://pubmed.ncbi.nlm.nih.gov/33470097/) · PMCID: [PMC8849040](https://pubmed.ncbi.nlm.nih.gov/PMC8849040/)
31. **SP3-FAIMS Chemoproteomics for High Coverage Profiling of the Human Cysteinome**
Tianyang Yan, Heta Desai, Lisa Boatner, Stephanie Yen, Jian Cao, Maria Palafox, Yasaman Jamialahmadi, Keriann Backus
American Chemical Society (ACS) (2020-12-28) <https://doi.org/gr3m63>
DOI: [10.26434/chemrxiv.13487364.v1](https://doi.org/10.26434/chemrxiv.13487364.v1)
32. **Expanding Proteome Coverage with Orthogonal-specificity α-Lytic Proteases**
Jesse G Meyer, Sangtae Kim, David A Maltby, Majid Ghassemian, Nuno Bandeira, Elizabeth A Komives
Molecular & Cellular Proteomics (2014-03) <https://doi.org/f5vgcg>
DOI: [10.1074/mcp.m113.034710](https://doi.org/10.1074/mcp.m113.034710) · PMID: [24425750](https://pubmed.ncbi.nlm.nih.gov/24425750/) · PMCID: [PMC3945911](https://pubmed.ncbi.nlm.nih.gov/PMC3945911/)
33. **XL-MS: Protein cross-linking coupled with mass spectrometry**
Andrew N Holding
Methods (2015-11) <https://doi.org/gftzb6>
DOI: [10.1016/j.ymeth.2015.06.010](https://doi.org/10.1016/j.ymeth.2015.06.010)
34. **To Cleave or Not To Cleave in XL-MS?**
B Steigenberger, P Albanese, AJR Heck, RA Scheltema
Journal of the American Society for Mass Spectrometry (2019-12-27) <https://doi.org/gr3m6s>
DOI: [10.1021/jasms.9b00085](https://doi.org/10.1021/jasms.9b00085)
35. **Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology**
Clinton Yu, Lan Huang
Analytical Chemistry (2017-11-21) <https://doi.org/gfvmcj>
DOI: [10.1021/acs.analchem.7b04431](https://doi.org/10.1021/acs.analchem.7b04431) · PMID: [29160693](https://pubmed.ncbi.nlm.nih.gov/29160693/) · PMCID: [PMC6022837](https://pubmed.ncbi.nlm.nih.gov/PMC6022837/)
36. **A New in Vivo Cross-linking Mass Spectrometry Platform to Define Protein-Protein Interactions in Living Cells**
Robyn M Kaake, Xiaorong Wang, Anthony Burke, Clinton Yu, Wynne Kandur, Yingying Yang, Eric J Novtisky, Tonya Second, Jicheng Duan, Athit Kao, ... Lan Huang
Molecular & Cellular Proteomics (2014-12) <https://doi.org/f6r7qw>
DOI: [10.1074/mcp.m114.042630](https://doi.org/10.1074/mcp.m114.042630) · PMID: [25253489](https://pubmed.ncbi.nlm.nih.gov/25253489/) · PMCID: [PMC4256503](https://pubmed.ncbi.nlm.nih.gov/PMC4256503/)
37. **Identification of cross-linked peptides from complex samples**
Bing Yang, Yan-Jie Wu, Ming Zhu, Sheng-Bo Fan, Jinzhong Lin, Kun Zhang, Shuang Li, Hao Chi, Yu-Xin Li, Hai-Feng Chen, ... Meng-Qiu Dong
Nature Methods (2012-07-08) <https://doi.org/f38rv2>
DOI: [10.1038/nmeth.2099](https://doi.org/10.1038/nmeth.2099)
38. **Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments**
Glenn R Masson, John E Burke, Natalie G Ahn, Ganesh S Anand, Christoph Borchers, Sébastien Brier, George M Bou-Assaf, John R Engen, SWalter Englander, Johan Faber, ... Kasper D Rand

39. **Hydrogen exchange mass spectrometry for studying protein structure and dynamics**
Lars Konermann, Jingxi Pan, Yu-Hong Liu
Chem. Soc. Rev. (2011) <https://doi.org/db943b>
DOI: [10.1039/c0cs00113a](https://doi.org/10.1039/c0cs00113a)
40. **Hydrogen exchange and structural dynamics of proteins and nucleic acids**
SWalter Englander, Neville R Kallenbach
Quarterly Reviews of Biophysics (1983-11) <https://doi.org/cxj4z7>
DOI: [10.1017/s0033583500005217](https://doi.org/10.1017/s0033583500005217)
41. **Differential hydrogen/deuterium exchange mass spectrometry analysis of protein-ligand interactions**
Michael J Chalmers, Scott A Busby, Bruce D Pascal, Graham M West, Patrick R Griffin
Expert Review of Proteomics (2011-02) <https://doi.org/btrkr9>
DOI: [10.1586/epr.10.109](https://doi.org/10.1586/epr.10.109) · PMID: [21329427](#) · PMCID: [PMC3113475](#)
42. **Hydrogen Exchange Mass Spectrometry: Are We Out of the Quicksand?**
Roxana E Iacob, John R Engen
Journal of the American Society for Mass Spectrometry (2012-04-03)
<https://doi.org/10.1007/s13361-012-0377-z>
DOI: [/10.1007/s13361-012-0377-z](https://doi.org/10.1007/s13361-012-0377-z)
43. **Replication in bioanalytical studies with HDX MS: aim as high as possible**
Jamie A Moroco, John R Engen
Bioanalysis (2015-05) <https://doi.org/gr3m65>
DOI: [10.4155/bio.15.46](https://doi.org/10.4155/bio.15.46) · PMID: [26039804](#) · PMCID: [PMC4837975](#)
44. **Development of a Peptide Probe for the Occurrence of Hydrogen (H_1/H_2) Scrambling upon Gas-Phase Fragmentation**
Kasper D Rand, Thomas JD Jørgensen
Analytical Chemistry (2007-10-13) <https://doi.org/c24d6z>
DOI: [10.1021/ac0710782](https://doi.org/10.1021/ac0710782)
45. **Hydroxyl Radical-Mediated Modification of Proteins as Probes for Structural Proteomics**
Guozhong Xu, Mark R Chance
Chemical Reviews (2007-08-01) <https://doi.org/c769n8>
DOI: [10.1021/cr0682047](https://doi.org/10.1021/cr0682047)
46. **Structural Mass Spectrometry of Proteins Using Hydroxyl Radical Based Protein Footprinting**
Liwen Wang, Mark R Chance
Analytical Chemistry (2011-08-01) <https://doi.org/fvrx3k>
DOI: [10.1021/ac200567u](https://doi.org/10.1021/ac200567u) · PMID: [21770468](#) · PMCID: [PMC3184339](#)
47. **Mapping nucleic acid structure by hydroxyl radical cleavage**
Thomas D Tullius, Jason A Greenbaum
Current Opinion in Chemical Biology (2005-04) <https://doi.org/cdcbtpp>
DOI: [10.1016/j.cbpa.2005.02.009](https://doi.org/10.1016/j.cbpa.2005.02.009)
48. **Probing DNA Structure with Hydroxyl Radicals**
Thomas D Tullius
Current Protocols in Nucleic Acid Chemistry (2001-12) <https://doi.org/bf493d>
DOI: [10.1002/0471142700.nc0607s02](https://doi.org/10.1002/0471142700.nc0607s02)

49. **Targeted DNA oxidation and trajectory of radical DNA using DFT based QM/MM dynamics**
Pradip K Biswas, Sandipan Chakraborty
Nucleic Acids Research (2019-02-18) <https://doi.org/gr3m6z>
DOI: [10.1093/nar/gkz089](https://doi.org/10.1093/nar/gkz089) · PMID: [30773597](https://pubmed.ncbi.nlm.nih.gov/30773597/) · PMCID: [PMC6451130](https://pubmed.ncbi.nlm.nih.gov/PMC6451130/)
50. **<scp>MS</scp> -Based Hydroxyl Radical Footprinting: Methodology and Application of Fast Photochemical Oxidation of Proteins (<scp>FPOP</scp>)**
Ben Niu, Michael L Gross
Mass Spectrometry-Based Chemical Proteomics (2019-07-26) <https://doi.org/gr3m6d>
DOI: [10.1002/9781118970195.ch15](https://doi.org/10.1002/9781118970195.ch15)
51. **Fenton and Fenton-like catalysts for electrochemical immunoassay: A mini review**
Jiejie Feng, Changshun Chu, Zhanfang Ma
Electrochemistry Communications (2021-04) <https://doi.org/gr3m6h>
DOI: [10.1016/j.elecom.2021.106970](https://doi.org/10.1016/j.elecom.2021.106970)
52. **Inactivation of MS2 coliphage in Fenton and Fenton-like systems: role of transition metals, hydrogen peroxide and sunlight**
Jessica I Nieto-Juarez, Katarzyna Pierzchła, Andrzej Sienkiewicz, Tamar Kohn
Environmental Science & Technology (2010-03-31) <https://doi.org/fmmv29>
DOI: [10.1021/es903739f](https://doi.org/10.1021/es903739f)
53. **The use of the methods of radiolysis to explore the mechanisms of free radical modifications in proteins**
Chantal Houée-Levin, Krzysztof Bobrowski
Journal of Proteomics (2013-10) <https://doi.org/f5kn8p>
DOI: [10.1016/j.jprot.2013.02.014](https://doi.org/10.1016/j.jprot.2013.02.014)
54. **High voltage electrical discharges followed by deep eutectic solvents extraction for the valorization of pomegranate seeds (*Punica granatum* L.)**
Ester Hernández-Corroto, Nadia Boussetta, María Luisa Marina, María Concepción García, Eugène Vorobiev
Innovative Food Science & Emerging Technologies (2022-07) <https://doi.org/gr3m6j>
DOI: [10.1016/j.ifset.2022.103055](https://doi.org/10.1016/j.ifset.2022.103055)
55. **Structural proteomics: Topology and relative accessibility of plant lipid droplet associated proteins**
Pascale Jolivet, Laure Aymé, Alexandre Giuliani, Frank Wien, Thierry Chardot, Yann Gohon
Journal of Proteomics (2017-10) <https://doi.org/gcnk7p>
DOI: [10.1016/j.jprot.2017.09.005](https://doi.org/10.1016/j.jprot.2017.09.005)
56. **Proteomic analysis of *Saccharomyces cerevisiae* response to plasma treatment**
Višnja Stulić, Tomislava Vukušić, Ana Butorac, Dean Popović, Zoran Herceg
International Journal of Food Microbiology (2019-03) <https://doi.org/gr3m6k>
DOI: [10.1016/j.ijfoodmicro.2018.12.017](https://doi.org/10.1016/j.ijfoodmicro.2018.12.017)
57. **Laser flash photolysis of hydrogen peroxide to oxidize protein solvent-accessible residues on the microsecond timescale**
David M Hambly, Michael L Gross
Journal of the American Society for Mass Spectrometry (2005-12-01) <https://doi.org/cjbbvw>
DOI: [10.1016/j.jasms.2005.09.008](https://doi.org/10.1016/j.jasms.2005.09.008)
58. **Fast photochemical oxidation of proteins (FPOP): A powerful mass spectrometry-based structural proteomics tool**
Danté T Johnson, Luciano H Di Stefano, Lisa M Jones

59. **Laser flash photochemical oxidation to locate heme binding and conformational changes in myoglobin**
David Hambly, Michael Gross
International Journal of Mass Spectrometry (2007-01) <https://doi.org/fvzqsg>
DOI: [10.1016/j.ijms.2006.08.018](https://doi.org/10.1016/j.ijms.2006.08.018)
60. **Modifications generated by fast photochemical oxidation of proteins reflect the native conformations of proteins**
Emily E Chea, Lisa M Jones
Protein Science (2018-04-14) <https://doi.org/gdpkmn>
DOI: [10.1002/pro.3408](https://doi.org/10.1002/pro.3408) · PMID: [29575296](#) · PMCID: [PMC5980583](#)
61. **Fast Photochemical Oxidation of Protein Footprints Faster than Protein Unfolding**
Brian C Gau, Joshua S Sharp, Don L Rempel, Michael L Gross
Analytical Chemistry (2009-07-23) <https://doi.org/djvkzx>
DOI: [10.1021/ac901054w](https://doi.org/10.1021/ac901054w) · PMID: [20337372](#) · PMCID: [PMC3164994](#)
62. **Protein painting reveals solvent-excluded drug targets hidden within native protein-protein interfaces**
Alessandra Luchini, Virginia Espina, Lance A Liotta
Nature Communications (2014-07-22) <https://doi.org/gr3m6t>
DOI: [10.1038/ncomms5413](https://doi.org/10.1038/ncomms5413) · PMID: [25048602](#) · PMCID: [PMC4109009](#)
63. **Protein painting, an optimized MS-based technique, reveals functionally relevant interfaces of the PD-1/PD-L1 complex and the YAP2/ZO-1 complex**
Amanda Haymond, Douglass Dey, Rachel Carter, Angela Dailing, Vaishnavi Nara, Pranavi Nara, Sravani Venkatayogi, Mikell Paige, Lance Liotta, Alessandra Luchini
Journal of Biological Chemistry (2019-07) <https://doi.org/ghmhrd>
DOI: [10.1074/jbc.ra118.007310](https://doi.org/10.1074/jbc.ra118.007310) · PMID: [31167787](#) · PMCID: [PMC6643031](#)
64. **Detecting Protein-Small Molecule Interactions Using Limited Proteolysis-Mass Spectrometry (LiP-MS)**
Monika Pepelnjak, Natalie de Souza, Paola Picotti
Trends in Biochemical Sciences (2020-10) <https://doi.org/gr3m6m>
DOI: [10.1016/j.tibs.2020.05.006](https://doi.org/10.1016/j.tibs.2020.05.006)
65. **Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry**
Simone Schopper, Abdullah Kahraman, Pascal Leuenberger, Yuehan Feng, Ilaria Piazza, Oliver Müller, Paul J Boersema, Paola Picotti
Nature Protocols (2017-10-26) <https://doi.org/gcf7zn>
DOI: [10.1038/nprot.2017.100](https://doi.org/10.1038/nprot.2017.100)
66. **Proteome-wide structural changes measured with limited proteolysis-mass spectrometry: an advanced protocol for high-throughput applications**
Liliana Malinovska, Valentina Cappelletti, Devon Kohler, Ilaria Piazza, Tsung-Heng Tsai, Monika Pepelnjak, Patrick Stalder, Christian Dörig, Fabian Sesterhenn, Franziska Elsässer, ... Paola Picotti
Nature Protocols (2022-12-16) <https://doi.org/gr3m6v>
DOI: [10.1038/s41596-022-00771-x](https://doi.org/10.1038/s41596-022-00771-x)

67. <https://www.eubopen.org/sites/www.eubopen.org/files/attachments/2022/LiP-MS%20protocol.pdf>
68. **Thermal proteome profiling for interrogating protein interactions**
André Mateus, Nils Kurzawa, Isabelle Becher, Sindhuja Sridharan, Dominic Helm, Frank Stein, Athanasios Typas, Mikhail M Savitski
Molecular Systems Biology (2020-03) <https://doi.org/gr3m62>
DOI: [10.1525/msb.20199232](https://doi.org/10.1525/msb.20199232) · PMID: [32133759](https://pubmed.ncbi.nlm.nih.gov/32133759/) · PMCID: [PMC7057112](https://pubmed.ncbi.nlm.nih.gov/PMC7057112/)
69. **Thermal proteome profiling: Insights into protein modifications, associations, and functions**
Cecile Le Sueur, Henrik M Hammarén, Sindhuja Sridharan, Mikhail M Savitski
Current Opinion in Chemical Biology (2022-12) <https://doi.org/grhp63>
DOI: [10.1016/j.cbpa.2022.102225](https://doi.org/10.1016/j.cbpa.2022.102225)
70. **Thermal Proteome Profiling Reveals the O-GlcNAc-Dependent Meltome**
Dustin T King, Jesús E Serrano-Negrón, Yanping Zhu, Christopher L Moore, Matthew D Shoulders, Leonard J Foster, David J Vocadlo
Journal of the American Chemical Society (2022-03-01) <https://doi.org/gqqm4n>
DOI: [10.1021/jacs.1c10621](https://doi.org/10.1021/jacs.1c10621) · PMID: [35230102](https://pubmed.ncbi.nlm.nih.gov/35230102/) · PMCID: [PMC8969899](https://pubmed.ncbi.nlm.nih.gov/PMC8969899/)
71. **Target Discovery Using Thermal Proteome Profiling**
Sindhuja Sridharan, Ina Günthner, Isabelle Becher, Mikhail Savitski, Marcus Bantscheff
Mass Spectrometry-Based Chemical Proteomics (2019-07-26) <https://doi.org/gr3m6c>
DOI: [10.1002/9781118970195.ch11](https://doi.org/10.1002/9781118970195.ch11)
72. **CETSA in integrated proteomics studies of cellular processes**
Nayana Prabhu, Lingyun Dai, Pär Nordlund
Current Opinion in Chemical Biology (2020-02) <https://doi.org/gh9nqs>
DOI: [10.1016/j.cbpa.2019.11.004](https://doi.org/10.1016/j.cbpa.2019.11.004)
73. **Quantitative proteomics: A strategic ally to map protein interaction networks**
Miguel Marcilla, Juan Pablo Albar
IUBMB Life (2012-12-24) <https://doi.org/f4gpwr>
DOI: [10.1002/iub.1081](https://doi.org/10.1002/iub.1081)
74. **Next-generation proteomics: towards an integrative view of proteome dynamics**
AFMaarten Altelaar, Javier Munoz, Albert JR Heck
Nature Reviews Genetics (2012-12-04) <https://doi.org/ghkhz2>
DOI: [10.1038/nrg3356](https://doi.org/10.1038/nrg3356)
75. **Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS**
Markku Varjosalo, Roberto Sacco, Alexey Stukalov, Audrey van Drogen, Melanie Planyavsky, Simon Hauri, Ruedi Aebersold, Keiryn L Bennett, Jacques Colinge, Matthias Gstaiger, Giulio Superti-Furga
Nature Methods (2013-03-03) <https://doi.org/f4t647>
DOI: [10.1038/nmeth.2400](https://doi.org/10.1038/nmeth.2400)
76. **The BioPlex Network: A Systematic Exploration of the Human Interactome**
Edward L Huttlin, Lily Ting, Raphael J Bruckner, Fana Gebreab, Melanie P Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, ... Steven P Gygi
Cell (2015-07) <https://doi.org/f7kc5h>
DOI: [10.1016/j.cell.2015.06.043](https://doi.org/10.1016/j.cell.2015.06.043) · PMID: [26186194](https://pubmed.ncbi.nlm.nih.gov/26186194/) · PMCID: [PMC4617211](https://pubmed.ncbi.nlm.nih.gov/PMC4617211/)

77. **Architecture of the human interactome defines protein communities and disease networks**
Edward L Huttlin, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P Gygi, Hannah Parzen, ... Wade Harper
Nature (2017-05) <https://doi.org/f96974>
DOI: [10.1038/nature22366](https://doi.org/10.1038/nature22366) · PMID: [28514442](https://pubmed.ncbi.nlm.nih.gov/28514442/) · PMCID: [PMC5531611](https://pubmed.ncbi.nlm.nih.gov/PMC5531611/)
78. **Dual proteome-scale networks reveal cell-specific remodeling of the human interactome**
Edward L Huttlin, Raphael J Bruckner, Jose Navarrete-Perea, Joe R Cannon, Kurt Baltier, Fana Gebreab, Melanie P Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, ... Steven P Gygi
Cell (2021-05) <https://doi.org/gkcn2g>
DOI: [10.1016/j.cell.2021.04.011](https://doi.org/10.1016/j.cell.2021.04.011) · PMID: [33961781](https://pubmed.ncbi.nlm.nih.gov/33961781/) · PMCID: [PMC8165030](https://pubmed.ncbi.nlm.nih.gov/PMC8165030/)
79. **Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy**
Jeffrey D Martell, Thomas J Deerinck, Yasemin Sancak, Thomas L Poulos, Vamsi K Mootha, Gina E Sosinsky, Mark H Ellisman, Alice Y Ting
Nature Biotechnology (2012-10-21) <https://doi.org/f4czsn>
DOI: [10.1038/nbt.2375](https://doi.org/10.1038/nbt.2375) · PMID: [23086203](https://pubmed.ncbi.nlm.nih.gov/23086203/) · PMCID: [PMC3699407](https://pubmed.ncbi.nlm.nih.gov/PMC3699407/)
80. **Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging**
Hyun-Woo Rhee, Peng Zou, Namrata D Udeshi, Jeffrey D Martell, Vamsi K Mootha, Steven A Carr, Alice Y Ting
Science (2013-03-15) <https://doi.org/f4qdfc>
DOI: [10.1126/science.1230593](https://doi.org/10.1126/science.1230593) · PMID: [23371551](https://pubmed.ncbi.nlm.nih.gov/23371551/) · PMCID: [PMC3916822](https://pubmed.ncbi.nlm.nih.gov/PMC3916822/)
81. **Proteomic Mapping of the Human Mitochondrial Intermembrane Space in Live Cells via Ratiometric APEX Tagging**
Victoria Hung, Peng Zou, Hyun-Woo Rhee, Namrata D Udeshi, Valentin Cracan, Tanya Svinkina, Steven A Carr, Vamsi K Mootha, Alice Y Ting
Molecular Cell (2014-07) <https://doi.org/f6d4pk>
DOI: [10.1016/j.molcel.2014.06.003](https://doi.org/10.1016/j.molcel.2014.06.003) · PMID: [25002142](https://pubmed.ncbi.nlm.nih.gov/25002142/) · PMCID: [PMC4743503](https://pubmed.ncbi.nlm.nih.gov/PMC4743503/)
82. **A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells**
Kyle J Roux, Dae In Kim, Manfred Raida, Brian Burke
Journal of Cell Biology (2012-03-12) <https://doi.org/f3wp8h>
DOI: [10.1083/jcb.201112098](https://doi.org/10.1083/jcb.201112098) · PMID: [22412018](https://pubmed.ncbi.nlm.nih.gov/22412018/) · PMCID: [PMC3308701](https://pubmed.ncbi.nlm.nih.gov/PMC3308701/)
83. **Proximity Dependent Biotinylation: Key Enzymes and Adaptation to Proteomics Approaches**
Payman Samavarchi-Tehrani, Reuben Samson, Anne-Claude Gingras
Molecular & Cellular Proteomics (2020-05) <https://doi.org/gr3r3r>
DOI: [10.1074/mcp.r120.001941](https://doi.org/10.1074/mcp.r120.001941) · PMID: [32127388](https://pubmed.ncbi.nlm.nih.gov/32127388/) · PMCID: [PMC7196579](https://pubmed.ncbi.nlm.nih.gov/PMC7196579/)
84. **Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID**
Renata Varnaitė, Stuart A MacNeill
PROTEOMICS (2016-07-27) <https://doi.org/f3qr9p>
DOI: [10.1002/pmic.201600123](https://doi.org/10.1002/pmic.201600123) · PMID: [27329485](https://pubmed.ncbi.nlm.nih.gov/27329485/) · PMCID: [PMC5053326](https://pubmed.ncbi.nlm.nih.gov/PMC5053326/)
85. **An improved smaller biotin ligase for BioID proximity labeling**
Dae In Kim, Samuel C Jensen, Kyle A Noble, Birendra KC, Kenneth H Roux, Khatereh Motamedchaboki, Kyle J Roux

86. **BioID: A Proximity-Dependent Labeling Approach in Proteomics Study**
Peipei Li, Yuan Meng, Li Wang, Li-jun Di
Functional Proteomics (2018-10-02) <https://doi.org/gr3m6f>
DOI: [10.1007/978-1-4939-8814-3_10](https://doi.org/10.1007/978-1-4939-8814-3_10)
87. **Guide to protein purification**
Richard R Burgess (editor)
Academic Press (2009)
ISBN: 9780123745361
88. **The protein protocols handbook**
John M Walker (editor)
Humana Press (2009)
ISBN: 9781597451987
89. **An overview of cell disruption methods for intracellular biomolecules recovery**
Tatiane Aparecida Gomes, Cristina Maria Zanette, Michele Rigon Spier
Preparative Biochemistry & Biotechnology (2020-02-19) <https://doi.org/gqpdr5>
DOI: [10.1080/10826068.2020.1728696](https://doi.org/10.1080/10826068.2020.1728696)
90. **Fast and Sensitive Total Protein and Peptide Assays for Proteomic Analysis**
Jacek R Wiśniewski, Fabienne Z Gaugaz
Analytical Chemistry (2015-04-09) <https://doi.org/f3nsk2>
DOI: [10.1021/ac504689z](https://doi.org/10.1021/ac504689z)
91. **Getting intimate with trypsin, the leading protease in proteomics**
Elien Vandermarliere, Michael Mueller, Lennart Martens
Mass Spectrometry Reviews (2013-06-15) <https://doi.org/gn64qb>
DOI: [10.1002/mas.21376](https://doi.org/10.1002/mas.21376)
92. **< i>In Silico</i>Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage**
Jesse G Meyer
ISRN Computational Biology (2014-04-22) <https://doi.org/gb6s2r>
DOI: [10.1155/2014/960902](https://doi.org/10.1155/2014/960902) · PMID: [30687733](#) · PMCID: [PMC6347401](#)
93. **Venomics and antivenomics of Indian spectacled cobra (Naja naja) from the Western Ghats**
Muralidharan Vanuopadath, Dileepkumar Raveendran, Bipin Gopalakrishnan Nair, Sudarslal Sadasivan Nair
Acta Tropica (2022-04) <https://doi.org/gpbzf7>
DOI: [10.1016/j.actatropica.2022.106324](https://doi.org/10.1016/j.actatropica.2022.106324)
94. **Sequencing-Grade < i>De novo</i> Analysis of MS/MS Triplets (CID/HCD/ETD) From Overlapping Peptides**
Adrian Guthals, Karl R Clauser, Ari M Frank, Nuno Bandeira
Journal of Proteome Research (2013-05-30) <https://doi.org/f47kqd>
DOI: [10.1021/pr400173d](https://doi.org/10.1021/pr400173d) · PMID: [23679345](#) · PMCID: [PMC4591044](#)
95. **Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data**
Rachel M Miller, Robert J Millikin, Connor V Hoffmann, Stefan K Solntsev, Gloria M Sheynkman, Michael R Shortreed, Lloyd M Smith
Journal of Proteome Research (2019-08-04) <https://doi.org/gpqmp2>

96. **Multi-protease Approach for the Improved Identification and Molecular Characterization of Small Proteins and Short Open Reading Frame-Encoded Peptides**
Philipp T Kaulich, Liam Cassidy, Jürgen Bartel, Ruth A Schmitz, Andreas Tholey
Journal of Proteome Research (2021-03-24) <https://doi.org/gpqmpz>
DOI: [10.1021/acs.jproteome.1c00115](https://doi.org/10.1021/acs.jproteome.1c00115)
97. **A Multiple Protease Strategy to Optimise the Shotgun Proteomics of Mature Medicinal Cannabis Buds**
Delphine Vincent, Vilnis Ezernieks, Simone Rochfort, German Spangenberg
International Journal of Molecular Sciences (2019-11-11) <https://doi.org/gpqmp3>
DOI: [10.3390/ijms20225630](https://doi.org/10.3390/ijms20225630) · PMID: [31717952](#) · PMCID: [PMC6888629](#)
98. **Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics**
Xiaofeng Guo, David C Trudgian, Andrew Lemoff, Sivaramakrishna Yadavalli, Hamid Mirzaei
Molecular & Cellular Proteomics (2014-06) <https://doi.org/f56bwx>
DOI: [10.1074/mcp.m113.035170](https://doi.org/10.1074/mcp.m113.035170) · PMID: [24696503](#) · PMCID: [PMC4047476](#)
99. **An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas**
Piero Giansanti, Thin Thin Aye, Henk van den Toorn, Mao Peng, Bas van Breukelen, Albert JR Heck
Cell Reports (2015-06) <https://doi.org/gpqmpx>
DOI: [10.1016/j.celrep.2015.05.029](https://doi.org/10.1016/j.celrep.2015.05.029)
100. **The Nobel Prize in Chemistry 1946**
NobelPrize.org
<https://www.nobelprize.org/prizes/chemistry/1946/speedread/>
101. **Online, High-Pressure Digestion System for Protein Characterization by Hydrogen/Deuterium Exchange and Mass Spectrometry**
Lisa M Jones, Hao Zhang, Ilan Vidavsky, Michael L Gross
Analytical Chemistry (2010-01-22) <https://doi.org/b993rm>
DOI: [10.1021/ac902477u](https://doi.org/10.1021/ac902477u) · PMID: [20095571](#) · PMCID: [PMC2826105](#)
102. **Hydrogen/deuterium exchange in mass spectrometry**
Yury Kostyukevich, Thamina Acter, Alexander Zhrebker, Arif Ahmed, Sunghwan Kim, Eugene Nikolaev
Mass Spectrometry Reviews (2018-03-30) <https://doi.org/gffzx8>
DOI: [10.1002/mas.21565](https://doi.org/10.1002/mas.21565)
103. **Proteinase K**
W Saenger
Handbook of Proteolytic Enzymes (2013) <https://doi.org/gkfkc2>
DOI: <https://doi.org/10.1016/b978-0-12-382219-2.00714-6> · ISBN: 9780123822192
104. **Insights into protein post-translational modification landscapes of individual human cells by trapped ion mobility time-of-flight mass spectrometry**
Benjamin C Orsburn, Yuting Yuan, Namandjé N Bumpus
Nature Communications (2022-11-25) <https://doi.org/grhnzi>
DOI: [10.1038/s41467-022-34919-w](https://doi.org/10.1038/s41467-022-34919-w) · PMID: [36433961](#) · PMCID: [PMC9700839](#)
105. **TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing**
Jiaming Li, Zhenying Cai, Ryan D Bomgarden, Ian Pike, Karsten Kuhn, John C Rogers, Thomas M Roberts, Steven P Gygi, Joao A Paulo

106. **Off-Line High-pH Reversed-Phase Fractionation for In-Depth Phosphoproteomics**
Tanveer S Batth, Chiara Francavilla, Jesper V Olsen
Journal of Proteome Research (2014-11-04) <https://doi.org/f72pvg>
DOI: [10.1021/pr500893m](https://doi.org/10.1021/pr500893m)
107. **Simultaneous Quantification of the Acetylome and Succinylome by ‘One-Pot’ Affinity Enrichment**
Nathan Basisty, Jesse G Meyer, Lei Wei, Bradford W Gibson, Birgit Schilling
PROTEOMICS (2018-08-19) <https://doi.org/gn4cmb>
DOI: [10.1002/pmic.201800123](https://doi.org/10.1002/pmic.201800123) · PMID: [30035354](https://pubmed.ncbi.nlm.nih.gov/30035354/) · PMCID: [PMC6175148](https://pubmed.ncbi.nlm.nih.gov/PMC6175148/)
108. **Bacterial cellulose nanofibers for albumin depletion from human serum**
Emel Tamahkar, Ceyhun Babaç, Tülin Kutsal, Erhan Pişkin, Adil Denizli
Process Biochemistry (2010-10) <https://doi.org/bzqhtf>
DOI: [10.1016/j.procbio.2010.07.007](https://doi.org/10.1016/j.procbio.2010.07.007)
109. **Electrostatic Repulsion Hydrophilic Interaction Chromatography for Isocratic Separation of Charged Solutes and Selective Isolation of Phosphopeptides**
Andrew J Alpert
Analytical Chemistry (2007-11-21) <https://doi.org/bt84c3>
DOI: [10.1021/ac070997p](https://doi.org/10.1021/ac070997p)
110. **Novel Application of Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) in Shotgun Proteomics: Comprehensive Profiling of Rat Kidney Proteome**
Piliang Hao, Tiannan Guo, Xin Li, Sunil S Adav, Jie Yang, Meng Wei, Siu Kwan Sze
Journal of Proteome Research (2010-06-10) <https://doi.org/bgkxkx>
DOI: [10.1021/pr100037h](https://doi.org/10.1021/pr100037h)
111. **The Nobel Prize in Chemistry 2002**
NobelPrize.org
<https://www.nobelprize.org/prizes/chemistry/2002/summary/>
112. **Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules**
Michael Karas, Doris Bachmann, Franz Hillenkamp
Analytical Chemistry (1985-12-01) <https://pubs.acs.org/doi/abs/10.1021/ac00291a042>
DOI: [10.1021/ac00291a042](https://doi.org/10.1021/ac00291a042)
113. **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**
Michael Karas, Franz Hillenkamp
Analytical Chemistry (1988-10-15) <https://doi.org/d577jp>
DOI: [10.1021/ac00171a028](https://doi.org/10.1021/ac00171a028)
114. **The Scientist :: Nobel Prize controversy** (2007-05-17)
<https://web.archive.org/web/20070517202246/http://cmbl.bjmu.edu.cn/news/0212/55.htm>
115. **α-Cyano-4-hydroxycinnamic acid, sinapinic acid, and ferulic acid as matrices and alkylating agents for matrix-assisted laser desorption/ionization time-of-flight mass spectrometric analysis of cysteine-containing peptides**
Hongmei Yang, Debin Wan, Fengrui Song, Zhiqiang Liu, Shuying Liu
Rapid communications in mass spectrometry: RCM (2013-06-30)
<https://pubmed.ncbi.nlm.nih.gov/23681820/>
DOI: [10.1002/rcm.6587](https://doi.org/10.1002/rcm.6587)

116. **The Desorption Process in MALDI**
Klaus Dreisewerd
Chemical Reviews (2003-01-24) <https://doi.org/cpzqmq>
DOI: [10.1021/cr010375i](https://doi.org/10.1021/cr010375i)
117. **Matrix Dependence of Metastable Fragmentation of Glycoproteins in MALDI TOF Mass Spectrometry**
Michael Karas, Ute Bahr, Kerstin Strupat, Franz Hillenkamp, Anthony Tsarbopoulos, Birendra N Pramanik
Analytical Chemistry (1995-02-01) <https://doi.org/b54gnt>
DOI: [10.1021/ac00099a029](https://doi.org/10.1021/ac00099a029)
118. **XX. < i>On the equilibrium of liquid conducting masses charged with electricity</i>**
Lord Rayleigh
The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science (1882-09) <https://doi.org/c6bp6h>
DOI: [10.1080/14786448208628425](https://doi.org/10.1080/14786448208628425)
119. **Capillary Liquid Chromatography/Mass Spectrometry for Peptide and Protein Characterization**
MT Davis, DC Stahl, KM Swiderek, TD Lee
Methods (1994-09) <https://doi.org/fmtw5k>
DOI: [10.1006/meth.1994.1031](https://doi.org/10.1006/meth.1994.1031)
120. **Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database**
John R Yates, Jimmy K Eng, Ashley L McCormack, David Schieltz
Analytical Chemistry (1995-04-15) <https://doi.org/dtcrm9>
DOI: [10.1021/ac00104a020](https://doi.org/10.1021/ac00104a020)
121. **Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ**
Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, Matthias Mann
Molecular & Cellular Proteomics (2014-09) <https://doi.org/f6hmm5>
DOI: [10.1074/mcp.m113.031591](https://doi.org/10.1074/mcp.m113.031591) · PMID: [24942700](https://pubmed.ncbi.nlm.nih.gov/24942700/) · PMCID: [PMC4159666](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC4159666/)
122. **The MaxQuant computational platform for mass spectrometry-based shotgun proteomics**
Stefka Tyanova, Tikira Temu, Juergen Cox
Nature Protocols (2016-10-27) <https://doi.org/f89fwn>
DOI: [10.1038/nprot.2016.136](https://doi.org/10.1038/nprot.2016.136)
123. **Advances in proteomics data analysis and display using an accurate mass and time tag approach**
Jennifer SD Zimmer, Matthew E Monroe, Wei-Jun Qian, Richard D Smith
Mass Spectrometry Reviews (2006) <https://doi.org/bqv76w>
DOI: [10.1002/mas.20071](https://doi.org/10.1002/mas.20071) · PMID: [16429408](https://pubmed.ncbi.nlm.nih.gov/16429408/) · PMCID: [PMC1829209](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC1829209/)
124. **A New Algorithm Using Cross-Assignment for Label-Free Quantitation with LC-LTQ-FT MS**
Victor P Andreev, Lingyun Li, Lei Cao, Ye Gu, Tomas Rejtar, Shiaw-Lin Wu, Barry L Karger
Journal of Proteome Research (2007-04-19) <https://doi.org/d3fshp>
DOI: [10.1021/pr0606880](https://doi.org/10.1021/pr0606880) · PMID: [17441747](https://pubmed.ncbi.nlm.nih.gov/17441747/) · PMCID: [PMC2563808](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC2563808/)
125. **IDEAL-Q, an Automated Tool for Label-free Quantitation Analysis Using an Efficient Peptide Alignment Approach and Spectral Data Validation**

Chih-Chiang Tsou, Chia-Feng Tsai, Ying-Hao Tsui, Putty-Reddy Sudhir, Yi-Ting Wang, Yu-Ju Chen, Jeou-Yuan Chen, Ting-Yi Sung, Wen-Lian Hsu

Molecular & Cellular Proteomics (2010-01) <https://doi.org/d9gxp>

DOI: [10.1074/mcp.m900177-mcp200](https://doi.org/10.1074/mcp.m900177-mcp200) · PMID: [19752006](https://pubmed.ncbi.nlm.nih.gov/19752006/) · PMCID: [PMC2808259](https://pubmed.ncbi.nlm.nih.gov/PMC2808259/)

126. **<i>SuperHirn</i> - a novel tool for high resolution LC-MS-based peptide/protein profiling**

Lukas N Mueller, Oliver Rinner, Alexander Schmidt, Simon Letarte, Bernd Bodenmiller, Mi-Youn Brusniak, Olga Vitek, Ruedi Aebersold, Markus Müller

PROTEOMICS (2007-08-28) <https://doi.org/c5zp29>

DOI: [10.1002/pmic.200700057](https://doi.org/10.1002/pmic.200700057)

127. **Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model**

Matthew Y Lim, João A Paulo, Steven P Gygi

Journal of Proteome Research (2019-09-23) <https://doi.org/ggd6nh>

DOI: [10.1021/acs.jproteome.9b00492](https://doi.org/10.1021/acs.jproteome.9b00492) · PMID: [31547658](https://pubmed.ncbi.nlm.nih.gov/31547658/) · PMCID: [PMC7346880](https://pubmed.ncbi.nlm.nih.gov/PMC7346880/)

128. **IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs**

Fengchao Yu, Sarah E Haynes, Alexey I Nesvizhskii

Molecular & Cellular Proteomics (2021) <https://doi.org/gqfrh3>

DOI: [10.1016/j.mcpro.2021.100077](https://doi.org/10.1016/j.mcpro.2021.100077) · PMID: [33813065](https://pubmed.ncbi.nlm.nih.gov/33813065/) · PMCID: [PMC8131922](https://pubmed.ncbi.nlm.nih.gov/PMC8131922/)

129. **Post Analysis Data Acquisition for the Iterative MS/MS Sampling of Proteomics Mixtures**

Michael R Hoopmann, Gennifer E Merrihew, Priska D von Haller, Michael J MacCoss

Journal of Proteome Research (2009-03-03) <https://doi.org/fgtwdr>

DOI: [10.1021/pr800828p](https://doi.org/10.1021/pr800828p) · PMID: [19256536](https://pubmed.ncbi.nlm.nih.gov/19256536/) · PMCID: [PMC2671646](https://pubmed.ncbi.nlm.nih.gov/PMC2671646/)

130. **The Implications of Proteolytic Background for Shotgun Proteomics**

Paola Picotti, Ruedi Aebersold, Bruno Domon

Molecular & Cellular Proteomics (2007-09) <https://doi.org/ffrsx7>

DOI: [10.1074/mcp.m700029-mcp200](https://doi.org/10.1074/mcp.m700029-mcp200)

131. **Large-scale characterization of HeLa cell nuclear phosphoproteins**

Sean A Beausoleil, Mark Jedrychowski, Daniel Schwartz, Joshua E Elias, Judit Villén, Jiaxu Li, Martin A Cohn, Lewis C Cantley, Steven P Gygi

Proceedings of the National Academy of Sciences (2004-08-09) <https://doi.org/cjmg8h>

DOI: [10.1073/pnas.0404720101](https://doi.org/10.1073/pnas.0404720101) · PMID: [15302935](https://pubmed.ncbi.nlm.nih.gov/15302935/) · PMCID: [PMC514446](https://pubmed.ncbi.nlm.nih.gov/PMC514446/)

132. **Evaluation of the utility of neutral-loss-dependent MS3 strategies in large-scale phosphorylation analysis**

Judit Villén, Sean A Beausoleil, Steven P Gygi

PROTEOMICS (2008-10-29) <https://doi.org/c52q48>

DOI: [10.1002/pmic.200800283](https://doi.org/10.1002/pmic.200800283) · PMID: [18972524](https://pubmed.ncbi.nlm.nih.gov/18972524/) · PMCID: [PMC2745099](https://pubmed.ncbi.nlm.nih.gov/PMC2745099/)

133. **Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of <i>N</i>-Linked Glycoproteins**

Charandeep Singh, Cleidiane G Zampronio, Andrew J Creese, Helen J Cooper

Journal of Proteome Research (2012-08-06) <https://doi.org/f36tfq>

DOI: [10.1021/pr300257c](https://doi.org/10.1021/pr300257c)

134. **Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques**

Jesse G Meyer, Birgit Schilling

135. **Review of software tools for design and analysis of large scale MRM proteomic datasets**
Christopher M Colangelo, Lisa Chung, Can Bruce, Kei-Hoi Cheung
Methods (2013-06) <https://doi.org/f449z8>
DOI: [10.1016/j.ymeth.2013.05.004](https://doi.org/10.1016/j.ymeth.2013.05.004) · PMID: [23702368](https://pubmed.ncbi.nlm.nih.gov/23702368/) · PMCID: [PMC3775261](https://pubmed.ncbi.nlm.nih.gov/PMC3775261/)
136. **Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards**
Eric Kuhn, Jiang Wu, Johann Karl, Hua Liao, Werner Zolg, Brad Guild
PROTEOMICS (2004-04) <https://doi.org/cxn7fw>
DOI: [10.1002/pmic.200300670](https://doi.org/10.1002/pmic.200300670)
137. **Mass Spectrometric Quantitation of Peptides and Proteins Using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA)**
NLeigh Anderson, Norman G Anderson, Lee R Haines, Darryl B Hardie, Robert W Olafson, Terry W Pearson
Journal of Proteome Research (2004-02-06) <https://doi.org/c494px>
DOI: [10.1021/pr034086h](https://doi.org/10.1021/pr034086h)
138. **Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins**
Leigh Anderson, Christie L Hunter
Molecular & Cellular Proteomics (2006-04) <https://doi.org/c8f5gz>
DOI: [10.1074/mcp.m500331-mcp200](https://doi.org/10.1074/mcp.m500331-mcp200)
139. **A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition**
Veronika Vidova, Zdenek Spacil
Analytica Chimica Acta (2017-04) <https://doi.org/f93hpx>
DOI: [10.1016/j.aca.2017.01.059](https://doi.org/10.1016/j.aca.2017.01.059)
140. **The current status of clinical proteomics and the use of MRM and MRM³ for biomarker validation**
Jérôme Lemoine, Tanguy Fortin, Arnaud Salvador, Aurore Jaffuel, Jean-Philippe Charrier, Geneviève Choquet-Kastylevsky
Expert Review of Molecular Diagnostics (2012-05) <https://doi.org/f337q3>
DOI: [10.1586/erm.12.32](https://doi.org/10.1586/erm.12.32)
141. **High Sensitivity Detection of Plasma Proteins by Multiple Reaction Monitoring of N-Glycosites**
Jianru Stahl-Zeng, Vinzenz Lange, Reto Ossola, Katrin Eckhardt, Wilhelm Krek, Ruedi Aebersold, Bruno Domon
Molecular & Cellular Proteomics (2007-10) <https://doi.org/bwgf39>
DOI: [10.1074/mcp.m700132-mcp200](https://doi.org/10.1074/mcp.m700132-mcp200)
142. **A database of mass spectrometric assays for the yeast proteome**
Paola Picotti, Henry Lam, David Campbell, Eric W Deutsch, Hamid Mirzaei, Jeff Ranish, Bruno Domon, Ruedi Aebersold
Nature Methods (2008-11) <https://doi.org/dvs3c7>
DOI: [10.1038/nmeth1108-913](https://doi.org/10.1038/nmeth1108-913) · PMID: [18974732](https://pubmed.ncbi.nlm.nih.gov/18974732/) · PMCID: [PMC2770732](https://pubmed.ncbi.nlm.nih.gov/PMC2770732/)

143. **Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics**
Amelia C Peterson, Jason D Russell, Derek J Bailey, Michael S Westphall, Joshua J Coon
Molecular & Cellular Proteomics (2012-11) <https://doi.org/f4gzw2>
DOI: [10.1074/mcp.o112.020131](https://doi.org/10.1074/mcp.o112.020131) · PMID: [22865924](https://pubmed.ncbi.nlm.nih.gov/22865924/) · PMCID: [PMC3494192](https://pubmed.ncbi.nlm.nih.gov/PMC3494192/)
144. **Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics**
Graziella E Ronsein, Nathalie Pamir, Priska D von Haller, Daniel S Kim, Michael N Oda, Gail P Jarvik, Tomas Vaisar, Jay W Heinecke
Journal of Proteomics (2015-01) <https://doi.org/f6wq8n>
DOI: [10.1016/j.jprot.2014.10.017](https://doi.org/10.1016/j.jprot.2014.10.017) · PMID: [25449833](https://pubmed.ncbi.nlm.nih.gov/25449833/) · PMCID: [PMC4259393](https://pubmed.ncbi.nlm.nih.gov/PMC4259393/)
145. **Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer**
Samuel Purvine, Jason-Thomas Eppel*, Eugene C Yi, David R Goodlett
PROTEOMICS (2003-06) <https://doi.org/c37bqm>
DOI: [10.1002/pmic.200300362](https://doi.org/10.1002/pmic.200300362)
146. **Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra**
John D Venable, Meng-Qiu Dong, James Wohlschlegel, Andrew Dillin, John R Yates III
Nature Methods (2004-09-29) <https://doi.org/dm8rm4>
DOI: [10.1038/nmeth705](https://doi.org/10.1038/nmeth705)
147. **UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation**
Robert S Plumb, Kelly A Johnson, Paul Rainville, Brian W Smith, Ian D Wilson, Jose M Castro-Perez, Jeremy K Nicholson
Rapid Communications in Mass Spectrometry (2006) <https://doi.org/c8gx39>
DOI: [10.1002/rcm.2550](https://doi.org/10.1002/rcm.2550)
148. **Precursor Acquisition Independent From Ion Count: How to Dive Deeper into the Proteomics Ocean**
Alexandre Panchaud, Alexander Scherl, Scott A Shaffer, Priska D von Haller, Hemantha D Kulasekara, Samuel I Miller, David R Goodlett
Analytical Chemistry (2009-07-02) <https://doi.org/dwbrjn>
DOI: [10.1021/ac900888s](https://doi.org/10.1021/ac900888s) · PMID: [19572557](https://pubmed.ncbi.nlm.nih.gov/19572557/) · PMCID: [PMC3086478](https://pubmed.ncbi.nlm.nih.gov/PMC3086478/)
149. **Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation**
Tamar Geiger, Juergen Cox, Matthias Mann
Molecular & Cellular Proteomics (2010-10) <https://doi.org/fqrbb3c>
DOI: [10.1074/mcp.m110.001537](https://doi.org/10.1074/mcp.m110.001537) · PMID: [20610777](https://pubmed.ncbi.nlm.nih.gov/20610777/) · PMCID: [PMC2953918](https://pubmed.ncbi.nlm.nih.gov/PMC2953918/)
150. **XDIA: improving on the label-free data-independent analysis**
Paulo C Carvalho, Xuemei Han, Tao Xu, Daniel Cociorva, Maria da Gloria Carvalho, Valmir C Barbosa, John R Yates III
Bioinformatics (2010-01-26) <https://doi.org/cn8x3r>
DOI: [10.1093/bioinformatics/btq031](https://doi.org/10.1093/bioinformatics/btq031) · PMID: [20106817](https://pubmed.ncbi.nlm.nih.gov/20106817/) · PMCID: [PMC2832823](https://pubmed.ncbi.nlm.nih.gov/PMC2832823/)
151. **ETISEQ – an algorithm for automated elution time ion sequencing of concurrently fragmented peptides for mass spectrometry-based proteomics**
Jason WH Wong, Alexander B Schwahn, Kevin M Downard
BMC Bioinformatics (2009-08-10) <https://doi.org/c3rshm>
DOI: [10.1186/1471-2105-10-244](https://doi.org/10.1186/1471-2105-10-244) · PMID: [19664259](https://pubmed.ncbi.nlm.nih.gov/19664259/) · PMCID: [PMC2731054](https://pubmed.ncbi.nlm.nih.gov/PMC2731054/)

152. **Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis**
Ludovic C Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, Ruedi Aebersold
Molecular & Cellular Proteomics (2012-06) <https://doi.org/fzqdc5>
DOI: [10.1074/mcp.o111.016717](https://doi.org/10.1074/mcp.o111.016717) · PMID: [22261725](https://pubmed.ncbi.nlm.nih.gov/22261725/) · PMCID: [PMC3433915](https://pubmed.ncbi.nlm.nih.gov/PMC3433915/)
153. **Multiplexed MS/MS for improved data-independent acquisition**
Jarrett D Egertson, Andreas Kuehn, Gennifer E Merrihew, Nicholas W Bateman, Brendan X MacLean, Ying S Ting, Jesse D Canterbury, Donald M Marsh, Markus Kellmann, Vlad Zabrouskov, ... Michael J MacCoss
Nature Methods (2013-06-23) <https://doi.org/gjqdct>
DOI: [10.1038/nmeth.2528](https://doi.org/10.1038/nmeth.2528) · PMID: [23793237](https://pubmed.ncbi.nlm.nih.gov/23793237/) · PMCID: [PMC3881977](https://pubmed.ncbi.nlm.nih.gov/PMC3881977/)
154. **Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results**
Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Yue Xuan, Julia Sondermann, Manuela Schmidt, David Gomez-Varela, Lukas Reiter
Molecular & Cellular Proteomics (2017-12) <https://doi.org/gcqvm4>
DOI: [10.1074/mcp.ra117.000314](https://doi.org/10.1074/mcp.ra117.000314) · PMID: [29070702](https://pubmed.ncbi.nlm.nih.gov/29070702/) · PMCID: [PMC5724188](https://pubmed.ncbi.nlm.nih.gov/PMC5724188/)
155. **diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition**
Florian Meier, Andreas-David Brunner, Max Frank, Annie Ha, Isabell Bludau, Eugenia Voytik, Stephanie Kaspar-Schoenfeld, Markus Lubeck, Oliver Raether, Nicolai Bache, ... Matthias Mann
Nature Methods (2020-11-30) <https://doi.org/gmhqdm>
DOI: [10.1038/s41592-020-00998-0](https://doi.org/10.1038/s41592-020-00998-0)
156. **Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry**
Yusuke Kawashima, Hirotaka Nagai, Ryo Konno, Masaki Ishikawa, Daisuke Nakajima, Hironori Sato, Ren Nakamura, Tomoyuki Furuyashiki, Osamu Ohara
Journal of Proteome Research (2022-05-06) <https://doi.org/gqgg5x>
DOI: [10.1021/acs.jproteome.2c00023](https://doi.org/10.1021/acs.jproteome.2c00023) · PMID: [35522919](https://pubmed.ncbi.nlm.nih.gov/35522919/) · PMCID: [PMC9171847](https://pubmed.ncbi.nlm.nih.gov/PMC9171847/)
157. **In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics**
Enrique Audain, Julian Uszkoreit, Timo Sachsenberg, Julianus Pfeuffer, Xiao Liang, Henning Hermjakob, Aniel Sanchez, Martin Eisenacher, Knut Reinert, David L Tabb, ... Yasset Perez-Riverol
Journal of Proteomics (2017-01) <https://doi.org/f9r8r6>
DOI: [10.1016/j.jprot.2016.08.002](https://doi.org/10.1016/j.jprot.2016.08.002)
158. **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**
Jürgen Cox, Matthias Mann
Nature Biotechnology (2008-11-30) <https://doi.org/crn24x>
DOI: [10.1038/nbt.1511](https://doi.org/10.1038/nbt.1511)
159. **MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics**
Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, Alexey I Nesvizhskii
Nature Methods (2017-04-10) <https://doi.org/f9z6p7>
DOI: [10.1038/nmeth.4256](https://doi.org/10.1038/nmeth.4256) · PMID: [28394336](https://pubmed.ncbi.nlm.nih.gov/28394336/) · PMCID: [PMC5409104](https://pubmed.ncbi.nlm.nih.gov/PMC5409104/)

160. **MS-GF+ makes progress towards a universal database search tool for proteomics**
Sangtae Kim, Pavel A Pevzner
Nature Communications (2014-10-31) <https://doi.org/ggkdq8>
DOI: [10.1038/ncomms6277](https://doi.org/10.1038/ncomms6277) · PMID: [25358478](https://pubmed.ncbi.nlm.nih.gov/25358478/) · PMCID: [PMC5036525](https://pubmed.ncbi.nlm.nih.gov/PMC5036525/)
161. **A method for reducing the time required to match protein sequences with tandem mass spectra**
Robertson Craig, Ronald C Beavis
Rapid Communications in Mass Spectrometry (2003) <https://doi.org/b7bgb9>
DOI: [10.1002/rcm.1198](https://doi.org/10.1002/rcm.1198)
162. **TANDEM: matching proteins with tandem mass spectra**
Robertson Craig, Ronald C Beavis
Bioinformatics (2004-02-19) <https://doi.org/cthw6n>
DOI: [10.1093/bioinformatics/bth092](https://doi.org/10.1093/bioinformatics/bth092)
163. **Skyline: an open source document editor for creating and analyzing targeted proteomics experiments**
Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, Michael J MacCoss
Bioinformatics (2010-02-09) <https://doi.org/bqx9rq>
DOI: [10.1093/bioinformatics/btq054](https://doi.org/10.1093/bioinformatics/btq054) · PMID: [2047306](https://pubmed.ncbi.nlm.nih.gov/2047306/) · PMCID: [PMC2844992](https://pubmed.ncbi.nlm.nih.gov/PMC2844992/)
164. **DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput**
Vadim Demichev, Christoph B Messner, Spyros I Vernalis, Kathryn S Lilley, Markus Ralser
Nature Methods (2019-11-25) <https://doi.org/gj9xgj>
DOI: [10.1038/s41592-019-0638-x](https://doi.org/10.1038/s41592-019-0638-x) · PMID: [31768060](https://pubmed.ncbi.nlm.nih.gov/31768060/) · PMCID: [PMC6949130](https://pubmed.ncbi.nlm.nih.gov/PMC6949130/)
165. **Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry**
Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, Ming Li
Nature Methods (2018-12-20) <https://doi.org/gftvmn>
DOI: [10.1038/s41592-018-0260-3](https://doi.org/10.1038/s41592-018-0260-3)
166. **Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues**
Roland Bruderer, Oliver M Bernhardt, Tejas Gandhi, Saša M Miladinović, Lin-Yang Cheng, Simon Messner, Tobias Ehrenberger, Vito Zanotelli, Yulia Butscheid, Claudia Escher, ... Lukas Reiter
Molecular & Cellular Proteomics (2015-05) <https://doi.org/f7b76h>
DOI: [10.1074/mcp.m114.044305](https://doi.org/10.1074/mcp.m114.044305) · PMID: [25724911](https://pubmed.ncbi.nlm.nih.gov/25724911/) · PMCID: [PMC4424408](https://pubmed.ncbi.nlm.nih.gov/PMC4424408/)
167. **PeptideShaker enables reanalysis of MS-derived proteomics data sets**
Marc Vaudel, Julia M Burkhart, René P Zahedi, Eystein Oveland, Frode S Berven, Albert Sickmann, Lennart Martens, Harald Barsnes
Nature Biotechnology (2015-01) <https://doi.org/ggkds8>
DOI: [10.1038/nbt.3109](https://doi.org/10.1038/nbt.3109)
168. **PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data**
Yehia Mokhtar Farag, Carlos Horro, Marc Vaudel, Harald Barsnes
Journal of Proteome Research (2021-10-28) <https://doi.org/gpdd85>
DOI: [10.1021/acs.jproteome.1c00678](https://doi.org/10.1021/acs.jproteome.1c00678) · PMID: [34709836](https://pubmed.ncbi.nlm.nih.gov/34709836/) · PMCID: [PMC8650087](https://pubmed.ncbi.nlm.nih.gov/PMC8650087/)

169. **mzML—a Community Standard for Mass Spectrometry Data**
Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpf, Steffen Neumann, Angel D Pizarro, ... Eric W Deutsch
Molecular & Cellular Proteomics (2011-01) <https://doi.org/dxkg99>
DOI: [10.1074/mcp.r110.000133](https://doi.org/10.1074/mcp.r110.000133) · PMID: [20716697](https://pubmed.ncbi.nlm.nih.gov/20716697/) · PMCID: [PMC3013463](https://pubmed.ncbi.nlm.nih.gov/PMC3013463/)
170. **File Formats Commonly Used in Mass Spectrometry Proteomics**
Eric W Deutsch
Molecular & Cellular Proteomics (2012-12) <https://doi.org/ggkdv>
DOI: [10.1074/mcp.r112.019695](https://doi.org/10.1074/mcp.r112.019695) · PMID: [22956731](https://pubmed.ncbi.nlm.nih.gov/22956731/) · PMCID: [PMC3518119](https://pubmed.ncbi.nlm.nih.gov/PMC3518119/)
171. **Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows**
Kenneth Verheggen, Helge Ræder, Frode S Berven, Lennart Martens, Harald Barsnes, Marc Vaudel
Mass Spectrometry Reviews (2017-09-13) <https://doi.org/gbwkmf>
DOI: [10.1002/mas.21543](https://doi.org/10.1002/mas.21543)
172. **Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases**
Lukas Käll, John D Storey, Michael J MacCoss, William Stafford Noble
Journal of Proteome Research (2007-12-08) <https://doi.org/fbxhxp>
DOI: [10.1021/pr700600n](https://doi.org/10.1021/pr700600n)
173. **False Discovery Rate Estimation in Proteomics**
Suruchi Aggarwal, Amit Kumar Yadav
Methods in Molecular Biology (2016) <https://doi.org/f79mzp>
DOI: [10.1007/978-1-4939-3106-4_7](https://doi.org/10.1007/978-1-4939-3106-4_7)
174. **Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach**
Lev I Levitsky, Mark V Ivanov, Anna A Lobas, Mikhail V Gorshkov
Journal of Proteome Research (2016-12-13) <https://doi.org/gqtfbj>
DOI: [10.1021/acs.jproteome.6b00144](https://doi.org/10.1021/acs.jproteome.6b00144)
175. **Quality Control in Proteomics**
PROTEOMICS
(2011-03) <https://doi.org/bzmh75>
DOI: [10.1002/pmic.201190020](https://doi.org/10.1002/pmic.201190020)
176. **Quality control in LC-MS/MS**
Thomas Köcher, Peter Pichler, Remco Swart, Karl Mechtler
PROTEOMICS (2011-02-07) <https://doi.org/c54ck5>
DOI: [10.1002/pmic.201000578](https://doi.org/10.1002/pmic.201000578)
177. **Quality control in mass spectrometry-based proteomics**
Wout Bittremieux, David L Tabb, Francis Impens, An Staes, Evy Timmerman, Lennart Martens, Kris Laukens
Mass Spectrometry Reviews (2017-09-07) <https://doi.org/gbs3vs>
DOI: [10.1002/mas.21544](https://doi.org/10.1002/mas.21544)
178. **AlphaPept, a modern and open framework for MS-based proteomics**
Maximilian T Strauss, Isabell Bludau, Wen-Feng Zeng, Eugenia Voytik, Constantin Ammar, Julia Schessner, Rajesh Ilango, Michelle Gill, Florian Meier, Sander Willemse, Matthias Mann

179. **rawDiag: An R Package Supporting Rational LC-MS Method Optimization for Bottom-up Proteomics**
Christian Trachsel, Christian Panse, Tobias Kockmann, Witold E Wolski, Jonas Grossmann, Ralph Schlapbach
Journal of Proteome Research (2018-07-06) <https://doi.org/gd39tz>
DOI: [10.1021/acs.jproteome.8b00173](https://doi.org/10.1021/acs.jproteome.8b00173)
180. **The rawrr R Package: Direct Access to Orbitrap Data and Beyond**
Tobias Kockmann, Christian Panse
Journal of Proteome Research (2021-03-09) <https://doi.org/gjgdxj>
DOI: [10.1021/acs.jproteome.0c00866](https://doi.org/10.1021/acs.jproteome.0c00866)
181. **RawBeans: A Simple, Vendor-Independent, Raw-Data Quality-Control Tool**
David Morgenstern, Rotem Barzilay, Yishai Levin
Journal of Proteome Research (2021-03-04) <https://doi.org/gmwh5f>
DOI: [10.1021/acs.jproteome.0c00956](https://doi.org/10.1021/acs.jproteome.0c00956) · PMID: [33657803](#) · PMCID: [PMC8041395](#)
182. **SIMPATIQCO: A Server-Based Software Suite Which Facilitates Monitoring the Time Course of LC-MS Performance Metrics on Orbitrap Instruments**
Peter Pichler, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, Christian G Huber, Christoph Stingl, Theo M Luider, Werner L Straube, Thomas Köcher, Karl Mechtler
Journal of Proteome Research (2012-10-22) <https://doi.org/f3sk5j>
DOI: [10.1021/pr300163u](https://doi.org/10.1021/pr300163u) · PMID: [23088386](#) · PMCID: [PMC3558011](#)
183. **Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data**
Bryan A Stanfill, Ernesto S Nakayasu, Lisa M Bramer, Allison M Thompson, Charles K Ansong, Therese R Clauss, Marina A Gritsenko, Matthew E Monroe, Ronald J Moore, Daniel J Orton, ... Thomas O Metz
Molecular & Cellular Proteomics (2018-09) <https://doi.org/gd7n2k>
DOI: [10.1074/mcp.ra118.000648](https://doi.org/10.1074/mcp.ra118.000648) · PMID: [29666158](#) · PMCID: [PMC6126382](#)
184. **SprayQc: A Real-Time LC-MS/MS Quality Monitoring System To Maximize Uptime Using Off the Shelf Components**
Richard A Scheltema, Matthias Mann
Journal of Proteome Research (2012-05-11) <https://doi.org/gqmt2m>
DOI: [10.1021/pr201219e](https://doi.org/10.1021/pr201219e)
185. **Metriculator: quality assessment for mass spectrometry-based proteomics**
Ryan M Taylor, Jamison Dance, Russ J Taylor, John T Prince
Bioinformatics (2013-09-02) <https://doi.org/f5gpzz>
DOI: [10.1093/bioinformatics/btt510](https://doi.org/10.1093/bioinformatics/btt510)
186. **OpenMS: a flexible open-source software platform for mass spectrometry data analysis**
Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, ... Oliver Kohlbacher
Nature Methods (2016-08-30) <https://doi.org/f82r32>
DOI: [10.1038/nmeth.3959](https://doi.org/10.1038/nmeth.3959)
187. **MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments**

Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, Olga Vitek
Bioinformatics (2014-05-02) <https://doi.org/f6j737>
DOI: [10.1093/bioinformatics/btu305](https://doi.org/10.1093/bioinformatics/btu305)

188. **MSstatsQC: Longitudinal System Suitability Monitoring and Quality Control for Targeted Proteomic Experiments**
Eralp Dogu, Sara Mohammad-Taheri, Susan E Abbatiello, Michael S Bereman, Brendan MacLean, Birgit Schilling, Olga Vitek
Molecular & Cellular Proteomics (2017-07) <https://doi.org/gbmgrh>
DOI: [10.1074/mcp.m116.064774](https://doi.org/10.1074/mcp.m116.064774) · PMID: [28483925](https://pubmed.ncbi.nlm.nih.gov/28483925/) · PMCID: [PMC5500765](https://pmcid.ncbi.nlm.nih.gov/PMC5500765/)
189. **Proteomics Quality Control: Quality Control Software for MaxQuant Results**
Chris Bielow, Guido Mastrobuoni, Stefan Kempa
Journal of Proteome Research (2015-12-28) <https://doi.org/f8c54w>
DOI: [10.1021/acs.jproteome.5b00780](https://doi.org/10.1021/acs.jproteome.5b00780)
190. **protti: an R package for comprehensive data analysis of peptide- and protein-centric bottom-up proteomics data**
Jan-Philipp Quast, Dina Schuster, Paola Picotti
Bioinformatics Advances (2021-12-10) <https://doi.org/gqmt2n>
DOI: [10.1093/bioadv/vbab041](https://doi.org/10.1093/bioadv/vbab041) · PMID: [36699412](https://pubmed.ncbi.nlm.nih.gov/36699412/) · PMCID: [PMC9710675](https://pmcid.ncbi.nlm.nih.gov/PMC9710675/)
191. **The NCBI Eukaryotic Genome Annotation Pipeline**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
192. **UniProt** <https://www.uniprot.org/help/biocuration>
193. https://www.psidev.info/sites/default/files/2018-03/MIAPE_MSI_1.1.pdf
194. **UniProt** https://www.uniprot.org/help/sequence_origin
195. **UniProt** https://www.uniprot.org/help/manual_curation
196. **fasta_utilities**
Phillip Wilmarth
(2023-08-01) https://github.com/pwilmart/fast_utilities
197. **Datasets**
NCBI
<https://www.ncbi.nlm.nih.gov/datasets>
198. **Eukaryotic genomes annotated at NCBI**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/
199. **Prokaryotic RefSeq Genomes** <https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>
200. **Eukaryotic RefSeq Genome Annotation Status**
https://www.ncbi.nlm.nih.gov/genome/annotation_euk/status/
201. **Training & Tutorials - Site Guide - NCBI** <https://www.ncbi.nlm.nih.gov/guide/training-tutorials/>
202. **Ensembl genome browser 110** <https://useast.ensembl.org/index.html>
203. **Ensembl Genomes** <http://ensemblgenomes.org/>
204. **cRAP protein sequences** <https://www.thegpm.org/crap/>

205. **Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics**
Ashley M Frankenfield, Jiawei Ni, Mustafa Ahmed, Ling Hao
Cold Spring Harbor Laboratory (2022-04-28) <https://doi.org/gp4xcp>
DOI: [10.1101/2022.04.27.489766](https://doi.org/10.1101/2022.04.27.489766)
206. **Closeness centrality | Network analysis of protein interaction data**
EMBL-EBI
<https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/building-and-analysing-ppins/topological-ppin-analysis/closeness-centrality/>
207. **Protein Bioinformatics**
Cathy H Wu, Cecilia N Arighi, Karen E Ross (editors)
Methods in Molecular Biology (2017) <https://doi.org/grmckg>
DOI: [10.1007/978-1-4939-6783-4](https://doi.org/10.1007/978-1-4939-6783-4)
208. **Recent advances in mass-spectrometry based proteomics software, tools and databases**
Ankit Halder, Ayushi Verma, Deetarup Biswas, Sanjeeva Srivastava
Drug Discovery Today: Technologies (2021-12) <https://doi.org/grmckh>
DOI: [10.1016/j.ddtec.2021.06.007](https://doi.org/10.1016/j.ddtec.2021.06.007)
209. **Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics**
Eric W Deutsch, Henry Lam, Ruedi Aebersold
Physiological Genomics (2008-03) <https://doi.org/d8zpvp>
DOI: [10.1152/physiolgenomics.00298.2007](https://doi.org/10.1152/physiolgenomics.00298.2007)
210. **Urgent need for consistent standards in functional enrichment analysis**
Kaumadi Wijesooriya, Sameer A Jadaan, Kaushalya L Perera, Tanuveer Kaur, Mark Ziemann
PLOS Computational Biology (2022-03-09) <https://doi.org/gpm77p>
DOI: [10.1371/journal.pcbi.1009935](https://doi.org/10.1371/journal.pcbi.1009935) · PMID: [35263338](#) · PMCID: [PMC8936487](#)
211. **On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P**
RA Fisher
Journal of the Royal Statistical Society (1922-01) <https://doi.org/frpswx>
DOI: [10.2307/2340521](https://doi.org/10.2307/2340521)
212. **Faculty Opinions recommendation of Gene co-expression analysis for functional classification and gene-disease predictions.**
Naoki Masuda
Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature (2021-11-13) <https://doi.org/grmckm>
DOI: [10.3410/f.727201931.793589848](https://doi.org/10.3410/f.727201931.793589848)
213. **Robust gene coexpression networks using signed distance correlation**
Javier Pardo-Díaz, Lyuba V Bozhilova, Mariano Beguerisse-Díaz, Philip S Poole, Charlotte M Deane, Gesine Reinert
Cold Spring Harbor Laboratory (2020-06-22) <https://doi.org/grmckj>
DOI: [10.1101/2020.06.21.163543](https://doi.org/10.1101/2020.06.21.163543)
214. **Community detection in graphs**
Santo Fortunato
Physics Reports (2010-02) <https://doi.org/b48x7f>
DOI: [10.1016/j.physrep.2009.11.002](https://doi.org/10.1016/j.physrep.2009.11.002)
215. **Fundamental differences**
Peter W Gunning, Edna C Hardeman

216. **Systematic detection of functional proteoform groups from bottom-up proteomic datasets**
Isabell Bludau, Max Frank, Christian Dörig, Yujia Cai, Moritz Heusel, George Rosenberger, Paola Picotti, Ben C Collins, Hannes Röst, Ruedi Aebersold
Cold Spring Harbor Laboratory (2020-12-22) <https://doi.org/grmckk>
DOI: [10.1101/2020.12.22.423928](https://doi.org/10.1101/2020.12.22.423928)
217. **Distance-based protein folding powered by deep learning**
Jinbo Xu
Proceedings of the National Academy of Sciences (2019-08-09) <https://doi.org/ggr3kr>
DOI: [10.1073/pnas.1821309116](https://doi.org/10.1073/pnas.1821309116) · PMID: [31399549](#) · PMCID: [PMC6708335](#)
218. **Highly accurate protein structure prediction with AlphaFold**
John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, ... Demis Hassabis
Nature (2021-07-15) <https://doi.org/gk7nfp>
DOI: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2) · PMID: [34265844](#) · PMCID: [PMC8371605](#)
219. **Molecular dynamics simulations: advances and applications**
Josep Gelpí, Adam Hospital, Ramón Goñi, Modesto Orozco
Advances and Applications in Bioinformatics and Chemistry (2015-11) <https://doi.org/gb9wxp>
DOI: [10.2147/aabc.s70333](https://doi.org/10.2147/aabc.s70333) · PMID: [26604800](#) · PMCID: [PMC4655909](#)
220. **Learning functional properties of proteins with language models**
Serbulent Unsal, Heval Atas, Muammer Albayrak, Kemal Turhan, Aybar C Acar, Tunca Doğan
Nature Machine Intelligence (2022-03-21) <https://doi.org/grckz2>
DOI: [10.1038/s42256-022-00457-9](https://doi.org/10.1038/s42256-022-00457-9)
221. **Learning the protein language: Evolution, structure, and function**
Tristan Bepler, Bonnie Berger
Cell Systems (2021-06) <https://doi.org/gqbqpp>
DOI: [10.1016/j.cels.2021.05.017](https://doi.org/10.1016/j.cels.2021.05.017) · PMID: [34139171](#) · PMCID: [PMC8238390](#)
222. **Computational prediction of protein-protein binding affinities**
Till Siebenmorgen, Martin Zacharias
WIREs Computational Molecular Science (2019-10-29) <https://doi.org/grmckf>
DOI: [10.1002/wcms.1448](https://doi.org/10.1002/wcms.1448)
223. **DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements: Table 1.**
Hao Luo, Yan Lin, Feng Gao, Chun-Ting Zhang, Ren Zhang
Nucleic Acids Research (2013-11-15) <https://doi.org/f5shv>
DOI: [10.1093/nar/gkt1131](https://doi.org/10.1093/nar/gkt1131) · PMID: [24243843](#) · PMCID: [PMC3965060](#)
224. **The phenotypic landscape of essential human genes**
Luke Funk, Kuan-Chung Su, Jimmy Ly, David Feldman, Avtar Singh, Brittanía Moodie, Paul C Blainey, Iain M Cheeseman
Cell (2022-11) <https://doi.org/gq68c5>
DOI: [10.1016/j.cell.2022.10.017](https://doi.org/10.1016/j.cell.2022.10.017) · PMID: [36347254](#) · PMCID: [PMC10482496](#)
225. **Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR**
Michael Boettcher, Michael T McManus
Molecular Cell (2015-05) <https://doi.org/f7dvmk>

DOI: [10.1016/j.molcel.2015.04.028](https://doi.org/10.1016/j.molcel.2015.04.028) · PMID: [26000843](#) · PMCID: [PMC4441801](#)

226. **A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity**
Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, Emmanuelle Charpentier
Science (2012-08-17) <https://doi.org/f22gdg>
DOI: [10.1126/science.1225829](https://doi.org/10.1126/science.1225829) · PMID: [22745249](#) · PMCID: [PMC6286148](#)
227. **A glance at genome editing with CRISPR-Cas9 technology**
Antara Barman, Bornali Deb, Supriyo Chakraborty
Current Genetics (2019-11-05) <https://doi.org/gjrcz5>
DOI: [10.1007/s00294-019-01040-3](https://doi.org/10.1007/s00294-019-01040-3)
228. **From DNA break repair pathways to CRISPR/Cas-mediated gene knock-in methods**
Mahere Rezazade Bazaz, Hesam Dehghani
Life Sciences (2022-04) <https://doi.org/gr5j2p>
DOI: [10.1016/j.lfs.2022.120409](https://doi.org/10.1016/j.lfs.2022.120409)
229. **Gene editing and its applications in biomedicine**
Guanglei Li, Xiangyang Li, Songkuan Zhuang, Liren Wang, Yifan Zhu, Yangcan Chen, Wen Sun, Zeguang Wu, Zhuo Zhou, Jia Chen, ... Wensheng Wei
Science China Life Sciences (2022-02-18) <https://doi.org/gp8msx>
DOI: [10.1007/s11427-021-2057-0](https://doi.org/10.1007/s11427-021-2057-0) · PMID: [35235150](#) · PMCID: [PMC8889061](#)
230. **A Review on Advanced CRISPR-Based Genome-Editing Tools: Base Editing and Prime Editing**
Ali Saber Sichani, Maryam Ranjbar, Maryam Baneshi, Farid Torabi Zadeh, Jafar Fallahi
Molecular Biotechnology (2022-12-22) <https://doi.org/gr5j2d>
DOI: [10.1007/s12033-022-00639-1](https://doi.org/10.1007/s12033-022-00639-1)
231. **Recent Advances in Genome-Engineering Strategies**
Michaela A Boti, Konstantina Athanasopoulou, Panagiotis G Adamopoulos, Diamantis C Sideris, Andreas Scorilas
Genes (2023-01-02) <https://doi.org/gr5j2w>
DOI: [10.3390/genes14010129](https://doi.org/10.3390/genes14010129) · PMID: [36672870](#) · PMCID: [PMC9859587](#)
232. **Latest Developed Strategies to Minimize the Off-Target Effects in CRISPR-Cas-Mediated Genome Editing**
Muhammad Naeem, Saman Majeed, Mubasher Zahir Hoque, Irshad Ahmad
Cells (2020-07-02) <https://doi.org/gmhjsc>
DOI: [10.3390/cells9071608](https://doi.org/10.3390/cells9071608) · PMID: [32630835](#) · PMCID: [PMC7407193](#)
233. **Targeted Protein Degradation Tools: Overview and Future Perspectives**
Yuri Prozzillo, Gaia Fattorini, Maria Virginia Santopietro, Luigi Suglia, Alessandra Ruggiero, Diego Ferreri, Giovanni Messina
Biology (2020-11-26) <https://doi.org/gmfb42>
DOI: [10.3390/biology9120421](https://doi.org/10.3390/biology9120421) · PMID: [33256092](#) · PMCID: [PMC7761331](#)
234. **Tet-On Systems For Doxycycline-inducible Gene Expression**
Atze T. Das, Liliane Tenenbaum, Ben Berkhouit
Current Gene Therapy (2016-06-16) <https://doi.org/f8sfzz>
DOI: [10.2174/1566523216666160524144041](https://doi.org/10.2174/1566523216666160524144041) · PMID: [27216914](#) · PMCID: [PMC5070417](#)
235. **How to Choose the Right Inducible Gene Expression System for Mammalian Studies?**
Kallunki, Barisic, Jäättelä, Liu
Cells (2019-07-30) <https://doi.org/gr5j2v>

236. **Rapid Creation of Stable Mammalian Cell Lines for Regulated Expression of Proteins Using the Gateway® Recombination Cloning Technology and Flp-In T-REx® Lines**
Jessica Spitzer, Markus Landthaler, Thomas Tuschl
Methods in Enzymology (2013) <https://doi.org/f5fkw5>
DOI: [10.1016/b978-0-12-418687-3.00008-2](https://doi.org/10.1016/b978-0-12-418687-3.00008-2)
237. **Using the Flp-In™ T-Rex™ System to Regulate GPCR Expression**
Richard J Ward, Elisa Alvarez-Curto, Graeme Milligan
Methods in Molecular Biology (2011) <https://doi.org/dsvq7p>
DOI: [10.1007/978-1-61779-126-0_2](https://doi.org/10.1007/978-1-61779-126-0_2)
238. **The Art of Validating Quantitative Proteomics Data**
David C Handler, Dana Pascovici, Mehdi Mirzaei, Vivek Gupta, Ghasem Hosseini Salekdeh, Paul A Haynes
PROTEOMICS (2018-11-25) <https://doi.org/gr5j2b>
DOI: [10.1002/pmic.201800222](https://doi.org/10.1002/pmic.201800222)
239. **Pulse-Chase Analysis for Studying Protein Synthesis and Maturation**
Susanne Fritzsche, Sebastian Springer
Current Protocols in Protein Science (2014-11) <https://doi.org/gr5jz9>
DOI: [10.1002/0471140864.ps3003s78](https://doi.org/10.1002/0471140864.ps3003s78)
240. **Degradation Parameters from Pulse-Chase Experiments**
Celine Sin, Davide Chiarugi, Angelo Valleriani
PLOS ONE (2016-05-16) <https://doi.org/gbmt2f>
DOI: [10.1371/journal.pone.0155028](https://doi.org/10.1371/journal.pone.0155028) · PMID: [27182698](https://pubmed.ncbi.nlm.nih.gov/27182698/) · PMCID: [PMC4868333](https://pubmed.ncbi.nlm.nih.gov/PMC4868333/)
241. **Mass Spectrometry Methods for Measuring Protein Stability**
Daniel D Vallejo, Carolina Rojas Ramírez, Kristine F Parson, Yilin Han, Varun V Gadkari, Brandon T Ruotolo
Chemical Reviews (2022-03-22) <https://doi.org/gr5j2q>
DOI: [10.1021/acs.chemrev.1c00857](https://doi.org/10.1021/acs.chemrev.1c00857) · PMID: [35316030](https://pubmed.ncbi.nlm.nih.gov/35316030/) · PMCID: [PMC9197173](https://pubmed.ncbi.nlm.nih.gov/PMC9197173/)
242. **New Developments in Quantitative Real-time Polymerase Chain Reaction Technology**
Current Issues in Molecular Biology
(2014) <https://doi.org/gr5j2t>
DOI: [10.21775/cimb.016.001](https://doi.org/10.21775/cimb.016.001)
243. **RNA-Seq: a revolutionary tool for transcriptomics**
Zhong Wang, Mark Gerstein, Michael Snyder
Nature Reviews Genetics (2009-01) <https://doi.org/crrz3v>
DOI: [10.1038/nrg2484](https://doi.org/10.1038/nrg2484) · PMID: [19015660](https://pubmed.ncbi.nlm.nih.gov/19015660/) · PMCID: [PMC2949280](https://pubmed.ncbi.nlm.nih.gov/PMC2949280/)
244. **Development and Applications of Superfolder and Split Fluorescent Protein Detection Systems in Biology**
Jean-Denis Pedelacq, Stéphanie Cabantous
International Journal of Molecular Sciences (2019-07-15) <https://doi.org/ggks55>
DOI: [10.3390/ijms20143479](https://doi.org/10.3390/ijms20143479) · PMID: [31311175](https://pubmed.ncbi.nlm.nih.gov/31311175/) · PMCID: [PMC6678664](https://pubmed.ncbi.nlm.nih.gov/PMC6678664/)
245. **Analysis of protein-protein proximities using the split-ubiquitin system**
N Lehming
Briefings in Functional Genomics and Proteomics (2002-01-01) <https://doi.org/b9btsv>
DOI: [10.1093/bfgp/1.3.230](https://doi.org/10.1093/bfgp/1.3.230)

246. **Split-luciferase complementary assay: applications, recent developments, and future perspectives**
Taha Azad, Amin Tashakor, Saman Hosseinkhani
Analytical and Bioanalytical Chemistry (2014-07-08) <https://doi.org/f6gkk9>
DOI: [10.1007/s00216-014-7980-8](https://doi.org/10.1007/s00216-014-7980-8)
247. **Tobacco Etch Virus protease: A shortcut across biotechnologies**
Francesca Cesaratto, Oscar R Burrone, Gianluca Petris
Journal of Biotechnology (2016-08) <https://doi.org/gr5j2k>
DOI: [10.1016/j.jbiotec.2016.06.012](https://doi.org/10.1016/j.jbiotec.2016.06.012)
248. **A split β-lactamase sensor for the detection of DNA modification by cisplatin and ruthenium-based chemotherapeutic drugs**
Samuel R Hinton, Elizabeth LS Corpuz, Karen L McFarlane Holman, Scott C Meyer
Journal of Inorganic Biochemistry (2022-11) <https://doi.org/gr5j2m>
DOI: [10.1016/j.jinorgbio.2022.111986](https://doi.org/10.1016/j.jinorgbio.2022.111986)
249. **Detecting Protein-Protein Interaction Based on Protein Fragment Complementation Assay**
Tianwen Wang, Ningning Yang, Chen Liang, Hongjv Xu, Yafei An, Sha Xiao, Mengyuan Zheng, Lu Liu, Gaozhan Wang, Lei Nie
Current Protein & Peptide Science (2020-08-21) <https://doi.org/gr5j2s>
DOI: [10.2174/1389203721666200213102829](https://doi.org/10.2174/1389203721666200213102829)
250. **Technologies for the identification and validation of protein-protein interactions**
Karoline Pichlerova, Jozef Hanes
General physiology and biophysics (2021) <https://doi.org/gr5j2z>
DOI: [10.4149/gpb_2021035](https://doi.org/10.4149/gpb_2021035)
251. **A novel genetic system to detect protein-protein interactions**
Stanley Fields, Ok-kyu Song
Nature (1989-07) <https://doi.org/c3ntbc>
DOI: [10.1038/340245a0](https://doi.org/10.1038/340245a0)
252. **Yeast Two-Hybrid, a Powerful Tool for Systems Biology**
Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, Uwe Schlattner
International Journal of Molecular Sciences (2009-06-18) <https://doi.org/d45mqr>
DOI: [10.3390/ijms10062763](https://doi.org/10.3390/ijms10062763) · PMID: [19582228](https://pubmed.ncbi.nlm.nih.gov/19582228/) · PMCID: [PMC2705515](https://pubmed.ncbi.nlm.nih.gov/PMC2705515/)
253. **The mammalian two-hybrid system as a powerful tool for high-throughput drug screening**
Daniela Patrício, Margarida Fardilha
Drug Discovery Today (2020-04) <https://doi.org/gr5j2j>
DOI: [10.1016/j.drudis.2020.01.022](https://doi.org/10.1016/j.drudis.2020.01.022)
254. **Isolation of Endogenously Assembled RNA-Protein Complexes Using Affinity Purification Based on Streptavidin Aptamer S1**
Yangchao Dong, Jing Yang, Wei Ye, Yuan Wang, Chuantao Ye, Daihui Weng, Huan Gao, Fanglin Zhang, Zhikai Xu, Yingfeng Lei
International Journal of Molecular Sciences (2015-09-16) <https://doi.org/f7xz9m>
DOI: [10.3390/ijms160922456](https://doi.org/10.3390/ijms160922456) · PMID: [26389898](https://pubmed.ncbi.nlm.nih.gov/26389898/) · PMCID: [PMC4613318](https://pubmed.ncbi.nlm.nih.gov/PMC4613318/)
255. **FRET as a biomolecular research tool — understanding its potential while avoiding pitfalls**
WRuss Algar, Niko Hildebrandt, Steven S Vogel, Igor L Medintz

256. **Split-protein systems: beyond binary protein–protein interactions**
Sujan S Shekhawat, Indraneel Ghosh
Current Opinion in Chemical Biology (2011-12) <https://doi.org/cpvhs8>
DOI: [10.1016/j.cbpa.2011.10.014](https://doi.org/10.1016/j.cbpa.2011.10.014) · PMID: [22070901](https://pubmed.ncbi.nlm.nih.gov/22070901/) · PMCID: [PMC3237955](https://pubmed.ncbi.nlm.nih.gov/PMC3237955/)
257. **Proximity ligation assay: an ultrasensitive method for protein quantification and its applications in pathogen detection**
Pengzhi Wang, Yi Yang, Tianqi Hong, Guoqiang Zhu
Applied Microbiology and Biotechnology (2021-01-11) <https://doi.org/gr5j2c>
DOI: [10.1007/s00253-020-11049-1](https://doi.org/10.1007/s00253-020-11049-1)
258. **Understanding FRET as a Research Tool for Cellular Studies**
Dilip Shrestha, Attila Jenei, Péter Nagy, György Vereb, János Szöllősi
International Journal of Molecular Sciences (2015-03-25) <https://doi.org/f7cbff>
DOI: [10.3390/ijms16046718](https://doi.org/10.3390/ijms16046718) · PMID: [25815593](https://pubmed.ncbi.nlm.nih.gov/25815593/) · PMCID: [PMC4424985](https://pubmed.ncbi.nlm.nih.gov/PMC4424985/)
259. **Illuminating insights into protein–protein interactions using bioluminescence resonance energy transfer (BRET)**
Kevin DG Pfleger, Karin A Eidne
Nature Methods (2006-02-17) <https://doi.org/fq5kjg>
DOI: [10.1038/nmeth841](https://doi.org/10.1038/nmeth841)
260. **Developments in FRET- and BRET-Based Biosensors**
Yuxin Wu, Tianyu Jiang
Micromachines (2022-10-20) <https://doi.org/gr5j2x>
DOI: [10.3390/mi13101789](https://doi.org/10.3390/mi13101789) · PMID: [36296141](https://pubmed.ncbi.nlm.nih.gov/36296141/) · PMCID: [PMC9610962](https://pubmed.ncbi.nlm.nih.gov/PMC9610962/)
261. **Fluorescence correlation spectroscopy as a tool for the study of the intracellular dynamics and biological fate of protein corona**
Marta Martinez-Moro, Desiré Di Silvio, Sergio E Moya
Biophysical Chemistry (2019-10) <https://doi.org/gr5j2f>
DOI: [10.1016/j.bpc.2019.106218](https://doi.org/10.1016/j.bpc.2019.106218)
262. **Proximity labeling methods for proteomic analysis of membrane proteins**
Francisco Zafra, Dolores Piniella
Journal of Proteomics (2022-07) <https://doi.org/gr5j2n>
DOI: [10.1016/j.jprot.2022.104620](https://doi.org/10.1016/j.jprot.2022.104620)
263. **Research Techniques Made Simple: Emerging Methods to Elucidate Protein Interactions through Spatial Proximity**
Yonglu Che, Paul A Khavari
Journal of Investigative Dermatology (2017-12) <https://doi.org/gcmmkp>
DOI: [10.1016/j.jid.2017.09.028](https://doi.org/10.1016/j.jid.2017.09.028)
264. **Analysis of Protein Interactions *in situ* by Proximity Ligation Assays**
Björn Koos, Linda Andersson, Carl-Magnus Claussion, Karin Grannas, Axel Klaesson, Gaëlle Cane, Ola Söderberg
Current Topics in Microbiology and Immunology (2013) <https://doi.org/f25nfj>
DOI: [10.1007/82_2013_334](https://doi.org/10.1007/82_2013_334)
265. **Cross-linking and mass spectrometry as a tool for studying the structural biology of ribonucleoproteins**
Chris P Sarnowski, Maria Bikaki, Alexander Leitner

Structure (2022-04) <https://doi.org/gr5jz6>

DOI: [10.1016/j.str.2022.03.003](https://doi.org/10.1016/j.str.2022.03.003)

266. **Protein–RNA interactions: from mass spectrometry to drug discovery**

Benjamin Steinmetz, Izabela Smok, Maria Bikaki, Alexander Leitner

Essays in Biochemistry (2023-03) <https://doi.org/gr5jz8>

DOI: [10.1042/ebc20220177](https://doi.org/10.1042/ebc20220177) · PMID: 36866608 · PMCID: [PMC10070478](#)

267. **Anatomy of a crosslinker**

Adam Belsom, Juri Rappaport

Current Opinion in Chemical Biology (2021-02) <https://doi.org/gk6ddr>

DOI: [10.1016/j.cbpa.2020.07.008](https://doi.org/10.1016/j.cbpa.2020.07.008)

268. **Analysis of Protein Interactions by Surface Plasmon Resonance**

Dennis G Drescher, Dakshnamurthy Selvakumar, Marian J Drescher

Protein-Protein Interactions in Human Disease, Part A (2018) <https://doi.org/gr5jz7>

DOI: [10.1016/bs.apcsb.2017.07.003](https://doi.org/10.1016/bs.apcsb.2017.07.003)

269. **Proteomic analyses using an accurate mass and time tag strategy**

Ljiljana Paša-Tolić, Christophe Masselon, Richard C Barry, Yufeng Shen, Richard D Smith

BioTechniques (2004-10) <https://doi.org/gqcns4>

DOI: [10.2144/04374rv01](https://doi.org/10.2144/04374rv01)

270. **Open collaborative writing with Manubot**

Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter

PLOS Computational Biology (2019-06-24) <https://doi.org/c7np>

DOI: [10.1371/journal.pcbi.1007128](https://doi.org/10.1371/journal.pcbi.1007128) · PMID: 31233491 · PMCID: [PMC6611653](#)