

VU LEARNING HUB

vulearninghub.xyz

HANDOUT

BIO601

Proteomics

BIO601 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

BIO601

Proteomics

HANDOUT

Proteomics is the large-scale study of proteins — their structures, functions, interactions, and modifications — as expressed by a genome, cell, tissue, or organism. This handout covers the full proteomics workflow: protein separation by electrophoresis and chromatography, mass spectrometry, protein identification, quantitative proteomics, post-translational modifications, and biological applications.

Table of Contents

Module 1	Introduction to Proteomics and the Proteome
Module 2	Protein Separation I — Gel Electrophoresis
Module 3	Protein Separation II — Chromatography
Module 4	Mass Spectrometry Fundamentals
Module 5	Peptide Mass Fingerprinting and Tandem MS
Module 6	Protein Identification and Sequence Databases
Module 7	Quantitative Proteomics
Module 8	Post-Translational Modifications
Module 9	Protein Structure, Interactions, and Applications

MODULE 1: INTRODUCTION TO PROTEOMICS AND THE PROTEOME

Learning Objectives

- Define proteomics and the proteome and explain their relationship to genomics
- Describe the dynamic nature of the proteome
- List the major goals and application areas of proteomics
- Compare genomics, transcriptomics, and proteomics

What is Proteomics?

The term **proteome** was coined by Marc Wilkins in 1994 and refers to the entire set of proteins expressed by a genome, cell, tissue, or organism at a given time under defined conditions. **Proteomics** is the large-scale, systematic study of the proteome — its composition, structure, function, interactions, and modifications.

Unlike the genome, which is essentially static and identical in nearly every cell of an organism, the proteome is highly **dynamic**: it changes with cell type, developmental stage, environmental conditions, and disease state. A single gene can give rise to many different protein products through alternative splicing and post-translational modifications, so the number of distinct proteins far exceeds the number of genes (humans have ~20,000 genes but potentially over a million proteoforms).

Why Proteins Matter More Than Genes Alone

Proteins are the actual workhorses of the cell: enzymes catalyze reactions, structural proteins give shape, receptors and transporters mediate communication and movement, and antibodies defend the body. Measuring mRNA levels (transcriptomics) does not reliably predict protein abundance because of differences in translation rates, protein stability, and degradation. Proteomics therefore provides a more direct picture of cellular function and phenotype.

This is why proteomics is central to **biomarker discovery** (proteins in blood or urine that signal disease), **drug target identification** (most drugs bind proteins), and understanding mechanisms of disease at the molecular level.

Branches of Proteomics

Expression proteomics compares protein abundance between samples (e.g., healthy vs. diseased tissue) to find differentially expressed proteins. **Structural proteomics** aims to determine the three-dimensional structures of all proteins. **Functional proteomics** studies protein functions, interactions, and participation in complexes and pathways. **Clinical proteomics** applies these tools to diagnosis, prognosis, and treatment monitoring.

Key Terms & Definitions

Proteome	The complete set of proteins expressed by a genome, cell, tissue, or organism at a particular time under specific conditions.
-----------------	---

Proteoform	A specific molecular form of a protein arising from genetic variation, alternative splicing, or post-translational modification.
Biomarker	A measurable protein (or other molecule) whose presence or level indicates a biological state such as disease.
Dynamic range	The ratio between the most and least abundant proteins in a sample; in plasma it spans 10-12 orders of magnitude, making analysis challenging.

Example: Proteome vs. Genome

The human genome contains about 3 billion base pairs and roughly 20,000 protein-coding genes. The human proteome, however, is estimated to contain over 1 million distinct proteoforms because one gene can produce multiple mRNA transcripts (alternative splicing) and each protein can carry different combinations of post-translational modifications such as phosphorylation, glycosylation, and ubiquitination.

Module Summary

- Proteomics studies the complete set of proteins (the proteome) expressed under defined conditions.
- The proteome is dynamic, unlike the static genome; protein levels cannot be inferred reliably from mRNA alone.
- Main branches: expression, structural, functional, and clinical proteomics.
- Key applications: biomarker discovery, drug targets, and disease mechanism studies.

Exam Pointers (Very Important)

- ★ Definition of proteome and who coined the term (Marc Wilkins, 1994).
- ★ Why the proteome is larger than the genome: alternative splicing + PTMs.
- ★ Differences between genomics, transcriptomics, and proteomics.

MODULE 2: PROTEIN SEPARATION I — GEL ELECTROPHORESIS

Learning Objectives

- Explain the principle of SDS-PAGE and isoelectric focusing
- Describe two-dimensional gel electrophoresis (2-DE) and its workflow
- Interpret a 2-D gel image and list its limitations

SDS-PAGE

SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) separates proteins primarily by molecular mass. SDS is an anionic detergent that denatures proteins and coats them with a uniform negative charge, so that migration through the polyacrylamide gel toward the positive electrode depends almost entirely on size: smaller proteins move faster through the gel matrix. A **stacking gel** (low acrylamide, pH 6.8) concentrates the sample into a thin band, and a **resolving gel** (higher acrylamide, pH 8.8) performs the separation.

After electrophoresis, proteins are visualized by staining — **Coomassie Brilliant Blue** (simple, moderately sensitive), **silver staining** (very sensitive but less compatible with mass spectrometry), or fluorescent dyes such as SYPRO Ruby. Molecular weight is estimated by comparing migration distance with a ladder of standard proteins.

Isoelectric Focusing (IEF)

Isoelectric focusing separates proteins by their **isoelectric point (pI)** — the pH at which a protein carries no net charge. A pH gradient is established in the gel (using carrier ampholytes or immobilized pH gradient strips), and each protein migrates until it reaches the pH equal to its pI, where it stops because its net charge becomes zero. IEF has very high resolving power and is the first dimension of 2-DE.

Two-Dimensional Gel Electrophoresis (2-DE)

2-DE combines IEF (first dimension, separation by charge/pI) with SDS-PAGE (second dimension, separation by mass), resolving up to several thousand protein spots on a single gel. The classic workflow is: protein extraction and solubilization, IEF on an IPG strip, equilibration of the strip in SDS buffer, second-dimension SDS-PAGE, staining, image analysis, spot picking, in-gel digestion, and mass spectrometric identification.

2-DE remains valuable because it displays intact proteins including isoforms and modified variants as distinct spots. Its limitations are well known: poor representation of very large, very small, very hydrophobic (membrane) proteins and extreme-pI proteins; limited dynamic range; gel-to-gel variability; and labor intensity. **DIGE** (difference gel electrophoresis) labels samples with different fluorescent cyanine dyes (Cy2, Cy3, Cy5) and runs them on the same gel with an internal standard, greatly improving quantitative comparison.

Key Terms & Definitions

SDS	Anionic detergent that denatures proteins and gives them uniform negative charge for size-based separation.
Isoelectric point (pI)	The pH at which a protein has zero net charge and stops migrating in IEF.
IPG strip	Immobilized pH gradient strip used for reproducible first-dimension isoelectric focusing.
DIGE	Difference gel electrophoresis; multiplexed fluorescent labeling for quantitative 2-DE comparison.

Example: Reading a 2-D Gel

On a 2-DE gel, the horizontal axis is pH (acidic proteins on the left, basic on the right) and the vertical axis is molecular mass (large proteins at the top, small at the bottom). A spot that appears in the disease sample but not in the control — or is much darker — represents a differentially expressed protein; it is excised, digested with trypsin, and identified by mass spectrometry.

Module Summary

- SDS-PAGE separates proteins by molecular mass; IEF separates by isoelectric point.
- 2-DE combines both dimensions and can resolve thousands of protein spots.
- Staining methods: Coomassie, silver, fluorescent dyes.
- DIGE enables multiplexed quantitative comparison on a single gel.

Exam Pointers (Very Important)

- ★ Principle of SDS-PAGE: why does separation depend on size?
- ★ What is pI and how does IEF work?
- ★ Steps of the 2-DE workflow in order.
- ★ Limitations of 2-DE and how DIGE addresses quantitation.

MODULE 3: PROTEIN SEPARATION II — CHROMATOGRAPHY

Learning Objectives

- Describe the main chromatographic modes used in proteomics
- Explain multidimensional protein identification technology (MudPIT)
- Compare gel-based and gel-free (shotgun) proteomics

Chromatographic Modes

Ion-exchange chromatography separates proteins or peptides by charge: cation exchangers bind positively charged molecules, anion exchangers bind negatively charged ones, and bound species are eluted with increasing salt concentration or a pH change. **Size-exclusion (gel filtration) chromatography** separates by hydrodynamic size — large molecules elute first because they are excluded from the pores of the stationary phase.

Reversed-phase chromatography (RPC), usually with a C18 stationary phase, separates peptides by hydrophobicity and is the standard final separation directly coupled to mass spectrometry (LC-MS). **Affinity chromatography** exploits specific biological interactions — for example, immobilized metal affinity chromatography (IMAC) or TiO₂ for phosphopeptides, lectins for glycopeptides, and antibodies for immunoaffinity enrichment.

HPLC and Nano-LC

High-performance liquid chromatography (HPLC) uses high pressure to push mobile phase through tightly packed columns, giving fast, high-resolution separations. In proteomics, **nano-LC** (flow rates of 200-500 nL/min, narrow-bore columns) is standard because the low flow dramatically improves electrospray ionization efficiency and sensitivity, allowing detection of attomole to femtomole peptide amounts.

Shotgun Proteomics and MudPIT

Shotgun (bottom-up) proteomics digests the entire protein mixture into peptides first and then separates and analyzes the peptides by LC-MS/MS — no gels are involved. Because a digested proteome yields hundreds of thousands of peptides, a single LC dimension is insufficient; **MudPIT** (multidimensional protein identification technology) couples strong cation exchange (SCX) with reversed-phase chromatography in a single column or online 2-D LC system, fractionating peptides by charge first and hydrophobicity second before they enter the mass spectrometer.

Gel-free shotgun workflows offer better coverage of membrane and extreme proteins, higher throughput, and easier automation than 2-DE, and are now the dominant approach in discovery proteomics. The trade-off is that connectivity between peptides and their parent proteoforms is inferred rather than directly observed (**protein inference problem**).

Key Terms & Definitions

Reversed-phase chromatography	Peptide separation by hydrophobicity on a C18 column; the standard LC mode coupled to MS.
Nano-LC	Low-flow liquid chromatography that maximizes electrospray sensitivity for proteomics.
MudPIT	Two-dimensional LC (SCX followed by reversed phase) coupled online to MS for shotgun proteomics.
Protein inference problem	The ambiguity of assigning shared peptides to specific parent proteins in bottom-up proteomics.

Module Summary

- Ion exchange (charge), size exclusion (size), reversed phase (hydrophobicity), and affinity (specific binding) are the key chromatographic modes.
- Nano-LC coupled to MS is the workhorse separation for modern proteomics.
- MudPIT uses two orthogonal LC dimensions to handle the complexity of digested proteomes.
- Shotgun proteomics digests first and analyzes peptides; it outperforms gels in coverage and throughput.

Exam Pointers (Very Important)

- ★ Match each chromatography mode to the property it separates by.
- ★ Why is nano-LC preferred for LC-MS proteomics?
- ★ What is MudPIT and why are two dimensions needed?
- ★ Gel-based vs. shotgun proteomics: advantages of each.

MODULE 4: MASS SPECTROMETRY FUNDAMENTALS

Learning Objectives

- Describe the components of a mass spectrometer
- Compare ESI and MALDI ionization
- Explain common mass analyzers and their performance metrics
- Define m/z , resolution, and mass accuracy

How a Mass Spectrometer Works

A mass spectrometer has three essential parts: an **ion source** that converts neutral molecules into gas-phase ions, a **mass analyzer** that separates ions by their mass-to-charge ratio (m/z), and a **detector** that records ion abundance. Because proteins and peptides are large, polar, and thermally fragile, they require **soft ionization** methods that produce intact molecular ions without excessive fragmentation.

Two soft-ionization breakthroughs made proteomics possible and earned the 2002 Nobel Prize in Chemistry (shared with Kurt Wüthrich): **electrospray ionization (ESI)** by John Fenn and **matrix-assisted laser desorption/ionization (MALDI)** by Koichi Tanaka.

ESI vs. MALDI

In **ESI**, the liquid sample is sprayed through a high-voltage capillary, forming charged droplets that evaporate and release multiply charged gas-phase ions. ESI couples naturally to liquid chromatography (LC-ESI-MS), produces multiply charged ions (so large peptides appear at measurable m/z), and is ideal for online LC-MS/MS. In **MALDI**, the sample is co-crystallized with an organic matrix, a laser pulse desorbs and ionizes the molecules, typically producing singly charged ions; it is fast, tolerant of salts, and suited to simple mixtures and imaging (MALDI imaging mass spectrometry).

Mass Analyzers

Time-of-flight (TOF) analyzers measure the time ions take to travel a field-free flight tube — lighter ions arrive first; they offer high speed and theoretically unlimited mass range. **Quadrupoles** use oscillating electric fields to filter ions of a selected m/z and are the basis of triple-quadrupole instruments for targeted quantitation (SRM/MRM). **Ion traps** (3-D and linear) trap ions and eject them selectively, enabling MS^n experiments. **Orbitrap** and **FT-ICR** analyzers measure ion oscillation frequencies and deliver the highest resolution and mass accuracy, making them the standard for discovery proteomics.

Key performance metrics: **resolution** (ability to distinguish close m/z values, $m/\Delta m$), **mass accuracy** (closeness of measured to true mass, in ppm), **sensitivity**, **dynamic range**, and **scan speed**.

Key Terms & Definitions

m/z	Mass-to-charge ratio; the quantity actually measured by a mass spectrometer.
-------------------------	--

Soft ionization	Ionization that produces intact molecular ions with minimal fragmentation (ESI, MALDI).
Resolution	$m/\Delta m$; the ability to separate two ions of similar m/z .
Tandem MS (MS/MS)	Two stages of mass analysis with fragmentation between them, used to sequence peptides.

Example: Reading an m/z Value

A peptide of monoisotopic mass 1500 Da carrying 2 protons appears at $m/z = (1500 + 2 \times 1.0073)/2 \approx 751.0$. The same peptide with 3 charges appears at $m/z \approx 501.0$. This is why ESI's multiply charged ions bring large peptides into the measurable m/z range of the analyzer.

Module Summary

- Mass spectrometers consist of ion source, mass analyzer, and detector.
- ESI (LC-compatible, multiply charged) and MALDI (laser/matrix, singly charged) are the soft-ionization workhorses.
- Common analyzers: TOF, quadrupole, ion trap, Orbitrap, FT-ICR.
- Performance is judged by resolution, mass accuracy, sensitivity, and speed.

Exam Pointers (Very Important)

- ★ Components of a mass spectrometer and function of each.
- ★ ESI vs. MALDI: principle and applications.
- ★ Which analyzer gives the highest resolution?
- ★ Calculate m/z for a given peptide mass and charge state.

MODULE 5: PEPTIDE MASS FINGERPRINTING AND TANDEM MS

Learning Objectives

- Explain peptide mass fingerprinting (PMF)
- Describe collision-induced dissociation and peptide fragmentation
- Interpret b- and y-ion series for peptide sequencing
- Outline the bottom-up proteomics workflow

The Bottom-Up Workflow

In **bottom-up (shotgun) proteomics**, proteins are digested into peptides — most commonly with **trypsin**, which cleaves after lysine and arginine (unless followed by proline) — and the peptides are separated by LC and analyzed by MS/MS. Trypsin is preferred because it generates peptides of ideal length (7-20 amino acids) with a basic C-terminal residue that promotes good fragmentation and ionization.

The alternative **top-down** approach analyzes intact proteins without digestion, preserving complete proteoform information (combinations of PTMs), but it is technically demanding and limited to simpler mixtures. **Middle-down** uses restricted digestion to produce longer peptides as a compromise.

Peptide Mass Fingerprinting

PMF measures the masses of all tryptic peptides from a single purified protein (e.g., one 2-D gel spot) in a single MS scan, typically by MALDI-TOF. The experimental peptide mass list is compared against theoretical digests of every protein in a database; the protein whose predicted peptide masses best match is the identification. PMF is fast and simple but requires a pure protein, a sequenced genome database, and good mass accuracy — it cannot identify mixtures or proteins with unexpected modifications.

Tandem MS and Peptide Sequencing

In **MS/MS**, a precursor peptide ion is selected in the first analyzer, fragmented — most often by **collision-induced dissociation (CID/HCD)**, where collisions with inert gas break the weakest bonds — and the fragment masses are measured in the second analyzer. Peptides fragment preferentially at the amide (peptide) bonds, producing **b-ions** (N-terminal fragments) and **y-ions** (C-terminal fragments).

Because consecutive y-ions (or b-ions) differ by the mass of one amino acid residue, the amino acid sequence can be read directly from the mass differences — this is **de novo sequencing**. In practice, database searching is more common: search engines (Mascot, SEQUEST, MaxQuant/Andromeda) compare each experimental MS/MS spectrum against theoretical spectra predicted from a protein database and score the matches, controlling false discoveries with a **target-decoy** strategy (typically 1% FDR).

Key Terms & Definitions

Trypsin	Protease cleaving after Lys/Arg; the standard digestion enzyme in proteomics.
PMF	Peptide mass fingerprinting; identification by matching a set of peptide masses to a database.
b-ions / y-ions	N-terminal and C-terminal peptide fragment ions used to read sequence from MS/MS spectra.
Target-decoy FDR	False discovery rate estimation by searching reversed/shuffled decoy sequences alongside real ones.

Example: Reading a y-ion Series

If consecutive y-ions appear at m/z 500, 613, 726, the mass differences are 113 and 113 Da. Residue mass 113 corresponds to leucine or isoleucine (isobaric — indistinguishable by mass alone). Reading all such steps spells out the peptide sequence from C- to N-terminus.

Module Summary

- Bottom-up proteomics digests proteins (usually with trypsin) and analyzes peptides by LC-MS/MS.
- PMF identifies pure proteins from peptide masses alone via database matching.
- CID/HCD fragmentation yields b- and y-ions whose mass differences reveal the amino acid sequence.
- Database search engines with target-decoy FDR control are the standard identification route.

Exam Pointers (Very Important)

- ★ Why is trypsin the preferred protease?
- ★ PMF: principle, requirements, limitations.
- ★ What are b- and y-ions? How is sequence read from them?
- ★ What is the target-decoy strategy and what FDR is standard?

MODULE 6: PROTEIN IDENTIFICATION AND SEQUENCE DATABASES

Learning Objectives

- Name the major protein sequence databases
- Explain how database search engines score peptide-spectrum matches
- Describe validation and false discovery control

Major Databases

UniProt is the central protein sequence resource: **UniProtKB/Swiss-Prot** contains manually curated, reviewed entries with functional annotation, while **UniProtKB/TrEMBL** holds automatically annotated translations. **NCBI RefSeq** provides curated reference sequences, and organism-specific resources (e.g., TAIR for Arabidopsis, SGD for yeast) add specialized annotation. Choosing the right database and taxonomy filter is critical — searching a human sample against the entire NCBI nr wastes time and inflates false positives.

Search Engines and Scoring

Search engines generate theoretical MS/MS spectra for every peptide in the database (considering the enzyme, allowed missed cleavages, and specified modifications) and score each experimental spectrum against candidates with matching precursor mass. **Mascot** uses a probability-based MOWSE score; **SEQUEST** uses cross-correlation (XCorr); **Andromeda** (in MaxQuant) uses a probabilistic score. Key search parameters include precursor and fragment mass tolerances, fixed modifications (e.g., carbamidomethylation of cysteine) and variable modifications (e.g., oxidation of methionine, phosphorylation).

Results are validated statistically: the **target-decoy** approach searches a concatenated database of real (target) and reversed (decoy) sequences; matches to decoys estimate the **false discovery rate**, and score thresholds are set to achieve typically 1% FDR at peptide and protein levels.

From Peptides to Proteins

Peptides shared between multiple proteins create the **protein inference problem**; the parsimony principle groups proteins sharing peptides and reports the minimal set explaining all observed peptides (protein groups). **Sequence coverage** — the percentage of the protein sequence covered by identified peptides — and the number of unique peptides are reported as confidence indicators. Contaminant databases (keratins, trypsin, BSA) are routinely included to flag laboratory contaminants.

Key Terms & Definitions

UniProtKB/Swiss-Prot	Manually curated, reviewed protein sequence database with functional annotation.
Mascot / SEQUEST / Andromeda	Common database search engines for peptide-spectrum matching.

FDR	False discovery rate; standard 1% cutoff for peptide and protein identifications.
Sequence coverage	Fraction of a protein's sequence represented by identified peptides.

Module Summary

- UniProt (Swiss-Prot reviewed, TrEMBL unreviewed) and RefSeq are the core databases.
- Search engines match experimental MS/MS spectra to theoretical spectra and score them.
- Target-decoy searching controls the false discovery rate, usually at 1%.
- Protein inference groups shared peptides by parsimony; contaminants are flagged explicitly.

Exam Pointers (Very Important)

- ★ Difference between Swiss-Prot and TrEMBL.
- ★ Name two search engines and their scoring basis.
- ★ How does target-decoy FDR control work?
- ★ What is the protein inference problem?

MODULE 7: QUANTITATIVE PROTEOMICS

Learning Objectives

- Compare label-free and label-based quantitation
- Explain SILAC, iTRAQ/TMT, and SRM/MS
- Describe experimental design considerations for quantitative studies

Label-Free Quantitation

Label-free quantitation compares separate LC-MS/MS runs without any isotopic labels. Two main strategies exist: **spectral counting** (more abundant proteins generate more MS/MS spectra) and **precursor ion intensity** (integrating the chromatographic peak area of each peptide, as in MaxQuant's LFQ algorithm). Label-free is simple, cheap, works on any sample including clinical material, and scales to many samples — but it demands highly reproducible chromatography and careful normalization, and missing values across runs are common.

Metabolic and Chemical Labeling

SILAC (stable isotope labeling by amino acids in cell culture) grows cells in media containing heavy ($^{13}\text{C}/^{15}\text{N}$ -labeled) lysine and arginine; heavy and light samples are mixed before processing, so quantitation from MS1 peak ratios is highly accurate — but it is limited to culturable cells. **iTRAQ** and **TMT** (tandem mass tags) chemically label peptides from up to 8 (iTRAQ) or 18 (TMTpro) samples with isobaric tags; samples are pooled and the reporter ions released during MS/MS give relative quantitation, enabling multiplexed comparison in a single run. A known caveat of isobaric tagging is **ratio compression** from co-isolated interfering precursors, mitigated by MS3-based quantitation (SPS-MS3).

Targeted Proteomics: SRM/MS and PRM

Selected reaction monitoring (SRM), also called **multiple reaction monitoring (MRM)**, on a triple quadrupole monitors predefined precursor-to-fragment transitions for target peptides, delivering the highest sensitivity, reproducibility, and quantitative accuracy — the gold standard for validating biomarker candidates. **Parallel reaction monitoring (PRM)** performs the same idea on high-resolution instruments (Orbitrap), recording all fragments for better specificity. **Data-independent acquisition (DIA/SWATH)** systematically fragments all precursors in wide m/z windows, combining comprehensive coverage with reproducible quantitation.

Key Terms & Definitions

SILAC	Metabolic heavy-isotope labeling in cell culture; accurate MS1-based quantitation.
iTRAQ / TMT	Isobaric chemical tags for multiplexed relative quantitation via MS/MS reporter ions.

SRM/MRM	Targeted quantitation monitoring specific precursor-fragment transitions on a triple quadrupole.
DIA/SWATH	Data-independent acquisition fragmenting all precursors in windows; comprehensive and reproducible.

Example: Choosing a Quantitation Strategy

To compare protein expression in treated vs. untreated cultured cells with maximum accuracy, use SILAC (mix early, minimal technical variation). To compare 10 clinical plasma samples where metabolic labeling is impossible, use TMT multiplexing or label-free LFQ. To validate three candidate biomarker proteins across 200 patient samples with clinical-grade precision, develop an SRM/MRM assay.

Module Summary

- Label-free (spectral counting, MS1 intensity) is simple and unlimited in samples but needs reproducible LC.
- SILAC labels metabolically; iTRAQ/TMT label chemically with isobaric tags for multiplexing.
- SRM/MRM is the most sensitive, targeted validation method; PRM and DIA are modern alternatives.
- Match the quantitation method to the sample type and study scale.

Exam Pointers (Very Important)

- ★ Label-free vs. labeled quantitation: principles and trade-offs.
- ★ How do SILAC and TMT differ? What is ratio compression?
- ★ What is SRM/MRM and when is it used?

MODULE 8: POST-TRANSLATIONAL MODIFICATIONS

Learning Objectives

- List major PTM types and their functional roles
- Explain how phosphorylation is enriched and analyzed
- Describe how MS localizes modification sites

Why PTMs Matter

Post-translational modifications are chemical changes to proteins after translation that regulate activity, localization, interactions, and degradation. Common PTMs include **phosphorylation** (Ser/Thr/Tyr; signaling), **glycosylation** (N- and O-linked; folding, recognition, stability), **ubiquitination** (degradation, signaling), **acetylation** and **methylation** (chromatin regulation), **proteolytic cleavage** (activation, e.g., zymogens), and **lipidation** (membrane anchoring). PTMs are sub-stoichiometric and dynamic, so they must be enriched before analysis.

Phosphoproteomics

Phosphopeptides are enriched with **IMAC** (immobilized metal affinity chromatography, Fe³⁺/Ga³⁺) or **TiO₂** chromatography, which selectively bind the phosphate group. In MS/MS, phosphoserine/phosphothreonine peptides characteristically lose phosphoric acid (neutral loss of 98 Da) under CID, a diagnostic signature; **ETD** (electron transfer dissociation) preserves the modification on the peptide backbone and is preferred for confident **site localization** — determining exactly which residue carries the phosphate when several candidates exist. Phosphoproteomics maps signaling networks, e.g., kinase cascades in cancer.

Other PTM Analyses

Glycoproteomics uses lectin affinity or hydrazide chemistry enrichment; glycans fragment characteristically (oxonium ions) and require specialized search tools. **Ubiquitination** leaves a diagnostic Gly-Gly remnant (114.04 Da) on lysine after tryptic digestion. Because modifications change peptide mass by defined **mass shifts** (phosphorylation +79.97 Da, oxidation +15.99 Da, acetylation +42.01 Da), database searches include them as variable modifications — but each added variable modification expands the search space and must be used judiciously.

Key Terms & Definitions

PTM	Post-translational modification; chemical alteration of a protein after translation.
IMAC / TiO₂	Enrichment methods selectively capturing phosphopeptides.
Site localization	Determining the exact modified residue within a peptide.
Mass shift	Defined mass added by a modification, e.g., +79.97 Da for phosphorylation.

Example: Phosphorylation Mass Shift

Addition of a phosphate group (HPO_3) increases peptide mass by 79.966 Da. A database search for phosphopeptides therefore looks for peptides whose measured mass equals the theoretical unmodified mass plus multiples of 79.966 Da, and MS/MS fragment ions confirm which serine, threonine, or tyrosine carries it.

Module Summary

- PTMs (phosphorylation, glycosylation, ubiquitination, acetylation, etc.) vastly expand proteome diversity and regulate function.
- Low-abundance PTMs require selective enrichment (IMAC/ TiO_2 for phospho, lectins for glyco).
- CID shows diagnostic neutral losses; ETD preserves labile PTMs for site localization.
- Modifications are detected via defined mass shifts in database searches.

Exam Pointers (Very Important)

- ★ Name four PTMs and one functional role of each.
- ★ How are phosphopeptides enriched? What is the 98 Da neutral loss?
- ★ Why is ETD useful for PTM analysis?
- ★ Mass shifts for phosphorylation and oxidation.

MODULE 9: PROTEIN STRUCTURE, INTERACTIONS, AND APPLICATIONS

Learning Objectives

- Describe structural proteomics approaches
- Explain methods for studying protein-protein interactions
- List clinical and biotechnological applications of proteomics

Structural Proteomics

Structural proteomics aims to determine 3-D structures on a proteome scale. **X-ray crystallography** gives atomic-resolution structures but requires well-diffracting crystals; **NMR spectroscopy** works in solution for smaller proteins and reveals dynamics; **cryo-electron microscopy (cryo-EM)** now resolves large complexes near-atomic resolution without crystals. Computationally, **homology modeling** predicts structure from related templates, and AI systems such as **AlphaFold** predict structures directly from sequence with remarkable accuracy, transforming structural proteomics. MS-based foot-printing methods — **hydrogen-deuterium exchange (HDX-MS)** and **cross-linking MS (XL-MS)** — probe conformation and interfaces in solution.

Protein-Protein Interactions

Proteins act in complexes and networks. **Yeast two-hybrid (Y2H)** screens binary interactions genetically; **affinity purification-MS (AP-MS)** isolates a tagged bait protein with its physiological partners from cell lysates for MS identification; **protein microarrays** test thousands of immobilized proteins against a labeled probe. Interaction data are assembled into networks analyzed with tools like STRING and Cytoscape to reveal functional modules and pathways.

Applications

Biomarker discovery: comparative plasma/tissue proteomics finds disease signatures (e.g., cancer, cardiovascular, neurodegenerative markers) for early diagnosis. **Drug discovery:** target identification and validation, mechanism-of-action studies, and toxicity profiling. **Systems biology:** integrating proteomics with genomics and metabolomics to model cellular networks. **Food and agriculture:** allergen detection, GMO characterization, crop stress responses. The **Human Proteome Project (HPP)** coordinates the global effort to map every human protein, including the challenging 'missing proteins' with no prior MS evidence.

Key Terms & Definitions

AlphaFold	AI system predicting protein 3-D structure from amino acid sequence.
AP-MS	Affinity purification coupled to MS for identifying protein interaction partners.
HDX-MS	Hydrogen-deuterium exchange MS probing protein conformation and dynamics.

Human Proteome Project

International initiative to map the entire human proteome.

Example: From Discovery to Biomarker

A typical biomarker pipeline: (1) discovery — compare plasma proteomes of 50 cancer patients vs. 50 controls by TMT or DIA, finding 30 differentially abundant proteins; (2) verification — measure the top candidates by SRM/MRM in 200 samples; (3) clinical validation — develop an ELISA/antibody assay for the best 2-3 markers and test in thousands of patients before regulatory approval.

Module Summary

- Structural proteomics uses crystallography, NMR, cryo-EM, HDX-MS, and AI prediction (AlphaFold).
- Y2H, AP-MS, and protein arrays map protein-protein interaction networks.
- Applications: biomarkers, drug targets, systems biology, food safety.
- The Human Proteome Project aims to catalog every human protein.

Exam Pointers (Very Important)

- ★ Compare X-ray crystallography, NMR, and cryo-EM.
- ★ How does AP-MS identify interaction partners?
- ★ Outline the biomarker discovery pipeline.
- ★ What is the Human Proteome Project?