Integrating Machine Learning and Physics-Based Models for Enhanced Computational Workflows in Mass Spectrometry

Juan Restrepo Lopez

München 2025

Integrating Machine Learning and Physics-Based Models for Enhanced Computational Workflows in Mass Spectrometry

Juan Restrepo Lopez

Dissertation
an der Fakültät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Juan Restrepo Lopez
aus Medellin, Kolumbien

München, den 17. Juli 2025

Erstgutachter: Prof. Dr. Petra Schwille
Zweitgutachter: Prof. Dr. Joachim Rädler
Tag der mündlichen Prüfung: 21. Oktober 2025

A mi padre, cuyo inocente amor por el conocimiento me impulsó a perseguir una carrera científica.

Zusammenfassung

Massenspektrometrie-basierte Proteomik ist eine grundlegende Technik der modernen Systembiologie und ermöglicht die Identifikation und Quantifizierung zahlreicher Proteine aus komplexen biologischen Proben. Trotz erheblicher Fortschritte bei Messinstrumenten und -strategien stellt die computergestützte Interpretation von Massenspektren weiterhin eine große Herausforderung dar.

Ich beginne diese Dissertation mit einer Einführung in die massenspektrometriebasierte Proteomik und erläutere den gesamten Workflow – von der experimentellen Durchführung bis zur Datenanalyse. Anschließend präsentiere ich zwei komplementäre Projekte, die durch die Integration von maschinellem Lernen und physikbasiertem Modellieren sowohl das Vertrauen in die Identifikation als auch die Peptidabdeckung in Proteomik-Workflows verbessern. Das erste Projekt trägt nicht nur zu einer höheren Vorhersagegenauigkeit bei, sondern liefert auch eine mechanistische Erklärung für ein neu beobachtetes Phänomen in Ionenmobilitätsexperimenten.

Das erste Projekt untersucht, wie strukturelle und physikalische Eigenschaften von Peptiden zur Verbesserung der Identifikation genutzt werden können. Hierbei liegt der Fokus auf Peptidkonformationen in der Gasphase. Für die Auswertung kombinieren wir Molekulardynamik-Simulationen, Berechnungen der Ionenmobilität und die Vorhersage von Kollisionsquerschnitten (engl. collision cross section, CCS) mit maschinellem Lernen. Dabei zeigt sich, dass bestimmte Peptide bimodale CCS-Verteilungen in der Gasphase aufweisen, was auf das Vorhandensein stabiler Konformere hinweist. Durch das Modellieren dieser Verteilungen und die Vorhersage von CCS-Werten aus Peptidsequenzen verbessern wir die Peptidzuordnung und verringern die Ambiguität bei der Identifikation. Diese integrative Strategie zeigt, dass selbst subtile physikalische Merkmale – kombiniert mit prädiktiven Modellen – zu bedeutenden Verbesserungen bei der Identifikation von Peptiden führen können.

Das zweite Projekt konzentriert sich auf die Entwicklung eines maschinellen Lernmodells zur Neugewichtung von Peptid-Spektrum-Zuordnungen (engl. peptide-spectrum matches, PSMs) in datenabhängigen Akquisitionsverfahren (engl. data-dependent acquisition, DDA). Die in gängigen Suchmaschinen verwendeten Algorithmen nutzen oft nicht das volle Potenzial der in modernen Spektren enthaltenen Informationen. Als Antwort darauf schlagen wir ein Modell vor, das PSMs auf Basis gelernter, unterscheidbarer Muster aus Spektrum-Sequenz-Paaren neu bewertet. Das Modell steigert signifikant die Genauigkeit der Identifikation in verschiedenen Datensätzen und gewährleistet eine gut kalibrierte Kontrolle der Fehlerentdeckungsrate (engl. false discovery rate, FDR). Dieser Ansatz zeigt, dass vertrauenswürdige Identifikationen allein auf Basis von Spektrum und Peptidsequenz – ohne zusätzliche Metadaten oder Feature-Engineering – möglich sind.

Zusammen unterstützen diese beiden Projekte eine übergeordnete Vision: die Integration datengetriebener Lernmethoden mit physikalisch sinnvollen Repräsentationen, um eine genauere und robustere Peptididentifikation in der massenspektrometrischen Proteomik zu ermöglichen. Durch die Weiterentwicklung sowohl des theoretischen als auch des praktischen Verständnisses des Problems der Peptididentifikation leistet diese Arbeit einen Beitrag zur Entwicklung zuverlässigerer und interpretierbarer rechnergestützter Auswertungspipelines in der Proteomik.

Abstract

Mass spectrometry-based proteomics is a cornerstone of modern systems biology, enabling the large-scale identification and quantification of proteins from complex biological samples. However, despite advances in instrumentation and data acquisition strategies, the computational interpretation of mass spectra remains a significant challenge.

I begin this thesis with an introduction to mass spectrometry-based proteomics, covering the workflow from experimental setup to data analysis. I then present two complementary projects that enhance identification confidence and peptide coverage in proteomics workflows by integrating machine learning with physics-based modeling. The first project not only contributes to improved predictive accuracy but also provides a mechanistic explanation for a newly observed phenomenon in ion mobility experiments.

The first project explores how structural and physical properties of peptides can be used to enhance identification, focusing on gas-phase peptide conformations. We combine molecular dynamics simulations, ion mobility calculations, and collision cross section prediction(CCS) with machine learning to show that certain peptides exhibit bimodal CCS distributions in the gas phase, reflecting the presence of stable conformers. By modeling these distributions and predicting CCS values from peptide sequences, we improve peptide matching and reduce ambiguity in identifications. This integrative strategy demonstrates that even subtle physical features—when combined with predictive modeling—can yield meaningful improvements in peptide discrimination.

The second project focuses on the development of a machine learning-based re-scoring framework for peptide-spectrum matches(PSMs) in data-dependent acquisition workflows. Existing scoring algorithms used by standard search engines often fail to fully leverage the rich features embedded in modern spectra. In response, we propose a model that re-evaluates the PSMs using learned discriminative patterns derived from spectrum-sequence pairs. The model significantly increases identification sensitivity across multiple datasets and maintains well-calibrated false discovery rate control. This approach demonstrates that high-confidence identifications can be obtained using only spectrum and peptide sequence as input, without requiring auxiliary metadata or feature engineering.

Together, these two projects support a broader vision: that integrating data-driven learning methods with physically meaningful representations enables more accurate and robust peptide identification in mass spectrometry proteomics. By advancing both the theoretical and practical understanding of the peptide identification problem, this work contributes to the development of more reliable and interpretable computational proteomics pipelines.

Publications and Manuscripts

Juan Restrepo†, Daniel Szoelloesi†, Tobias Kiermeyer, Christoph Wichmann, Helmut Grubmueller, Juergen Cox. *Bimodal peptide collision cross section distribution reflects two stable conformations in the gas phase.* Under review, Nature Communications (see chapter 4)

Maximilien Burq†, Dejan Stepec†, **Juan Restrepo**†, Jure Zbontar, Shamil Urazbakhtin, Bryan Crampton, Shivani Tiwary, Rehan Chinoy, Melissa Miao, Juergen Cox, Peter Cimermancic. *Back to Basics: Spectrum and Peptide Sequence are Sufficient for Top-tier Mass Spectrometry Proteomics Identification.* Under review, Nature Methods (see chapter 5)

† These authors contributed equally to this work

Contents

Part I

Introduction

1
Mass-spectrometry-based proteomics

1.1 Shotgun Proteomics: A modern Mass Spectrometry-based workflow for proteomics

Proteomics is the large-scale study of proteins, the essential functional molecules within living organisms. Proteins perform a vast array of biological functions, acting as enzymes, structural components, signaling molecules, and molecular machines that sustain cellular life. Proteomics seeks to characterize the full set of proteins, collectively known as the proteome, that are expressed by a genome, cell, tissue, or organism at a specific time under defined conditions.

Mass Spectrometry (MS) has emerged as a central analytical platform for proteomics due to its unparalleled sensitivity, dynamic range, and ability to analyze complex protein mixtures. The versatility of MS supports diverse applications such as biomarker discovery, drug development, and systems biology [1]. MS-based proteomics has enabled researchers to access previously unattainable layers of biological regulation, including post-translational modifications and protein-protein interactions. Among the many approaches, shotgun proteomics has become a cornerstone technique, providing a high-throughput and unbiased strategy for protein identification and quantification.

In this section, I will progress from simple to advanced concepts, providing both intuitive and technical insight into modern MS-based proteomics workflows. We begin by walking through a simplified proteomics workflow. Then we try to demystify the operating principles of a basic mass spectrometer. Finally, we present a comprehensive view of a state-of-the-art proteomics pipeline, which integrates orthogonal separation technologies and tandem mass spectrometry to tackle the challenges of biological complexity.

1.1.1 A Minimal Proteomics Workflow: From Sample to Spectrum

The goal of this section is to trace the path from a biological sample to the generation of a mass spectrum to be used later for identification. Although this minimal workflow omits many refinements found in modern systems, it captures the essential stages that make proteomics with mass spectrometry possible.

In this conceptual setup, a protein sample undergoes three major stages: sample extraction, enzymatic digestion, and mass spectrometry analysis; see Fig.1.1. First,

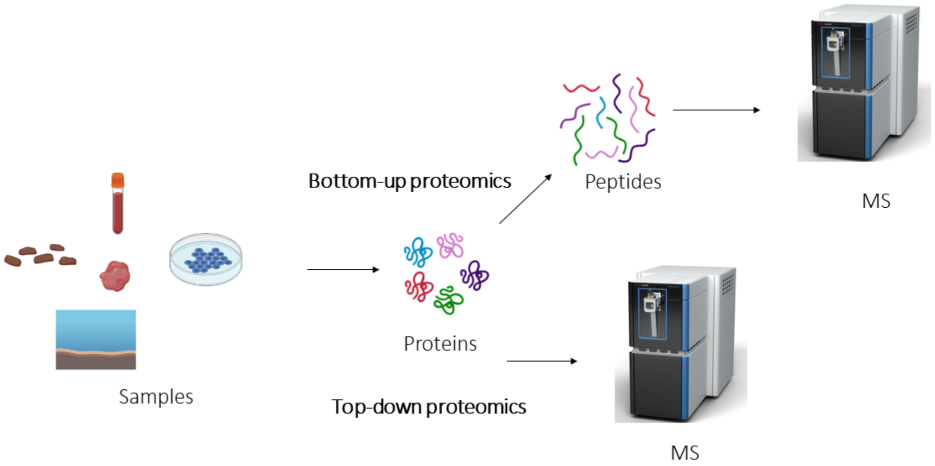


Figure 1.1. Minimal Proteomics Workflow. Created with BioRender.com.

proteins are extracted from a complex biological matrix, such as a cell lysate or a blood sample, yielding a heterogeneous mixture of soluble proteins. In the second stage, proteins are enzymatically digested - often with trypsin - into smaller peptide fragments of smaller dynamic range. This approach is known as bottom-up proteomics. Alternatively, in top-down proteomics, intact proteins are directly analyzed without prior digestion [2]. In the third and last stage, the resulting peptides (or intact proteins) are ionized and analyzed with the mass spectrometer. At this point, we treat the mass spectrometer as a functional black box that outputs a mass spectrum: a set of peaks representing the detected peptide ions.

Understanding how peptides are converted into mass spectra requires a deeper look at the internal mechanisms of the mass spectrometer. Although modern spectrometers are complex and highly engineered, their core functionality relies on well-established physical principles governing the behavior of ions in electric and magnetic fields.

1.1.2 A Simplified Mass Spectrometer: An Introduction to Physical Principles

A mass spectrometer is an advanced analytical instrument that measures the mass-to-charge ratio (m/z) of ions. To understand how it works, let us use a simple device that is no longer used in modern laboratories; see Fig.1.2. The process begins by injecting the sample that comes in an aqueous or organic solution into the mass spectrometer. There, it is immediately vaporized by a heater and then bombarded by high-energy electrons to create positive ions out of the neutral molecules. This ionization method is called electron impact ionization[3].

Once ionized, the ions are accelerated through a potential difference in an electric field, which imparts kinetic energy to the ions. From the law of conservation of energy, we see that the energy of the electric potential will transform into the kinetic energy of the ions and they will gain a speed given by:

$$K_E = U_E \implies \frac{1}{2}mv^2 = qV \implies v^2 = \frac{2qV}{m} \tag{1.1}$$

where K_E is the kinetic energy of the ion at the end of the acceleration region, U_E is the potential energy at the beginning of the acceleration region, m is the mass of the ion, v is its velocity, q is the ion's charge, and V is the potential difference across the electrodes. As a result, the velocity is determined by the ion's mass and charge, as lighter ions with the same charge will move faster than heavier ions.

After acceleration, the ions pass through a magnetic field where they experience deflection. The amount of deflection is influenced by the ion's mass-to-charge ratio. This deflection follows the Lorentz force law, which describes the force acting on a charged particle moving through a magnetic field. The equation for the force experienced by the ion is:

$$\vec{F} = q\vec{v} \times \vec{B} \tag{1.2}$$

where $\vec{F}$ is the force on the ion, q is its charge, $\vec{v}$ is the ion's velocity, $\vec{B}$ is the magnetic field, and θ is the angle between the velocity vector and the magnetic field.

If we take the magnetic field to be constant along the z axis and to be perpendicular to the velocity field we get that

$$F_x = qv_yB_z \implies m\ddot{x} = q\dot{y}B_z, \quad F_y = -qv_xB_z \implies m\ddot{y} = -q\dot{x}B_z, \quad F_z = 0 \tag{1.3}$$

The interesting dynamics takes place on the xy plane. The relevant equations are a system of coupled ordinary differential equations that have as a solution

$$x(t) = R\cos\left(\omega t + \phi\right) + B, \quad y(t) = R\sin\left(\omega t + \phi\right) + C \tag{1.4}$$

where $\omega = \frac{qB_z}{m}$ and B, C, ϕ and R are constant to be determined by the initial conditions. These equations represents a circular motion of radius R, angular frequency ω and centered in a point dependent on C and B.

In the simplest case where we start from the origin with velocity only along the x-axis we get that

$$R = \frac{mv}{qB} \tag{1.5}$$

Thus, ions with a higher mass will have a larger radius of curvature than lighter ions, and ions with different m/z ratios will follow different trajectories, leading to their separation.

Finally, the separated ions are detected typically by converting the ion current into an electric current proportional to it.

The mass spectrum, which is a plot of ion signal intensity versus m/z, provides detailed information about the sample's ionic composition. The position of each peak along the x-axis corresponds to the m/z ratio of the ion, and the height of the peak represents the relative abundance of that ion in the sample.

Although this description refers to a simpler earlier mass spectrometer, the physical principles of the more modern instruments remain the same. Today's mass spectrometers differ in the ionization, acceleration, deflection, and detection

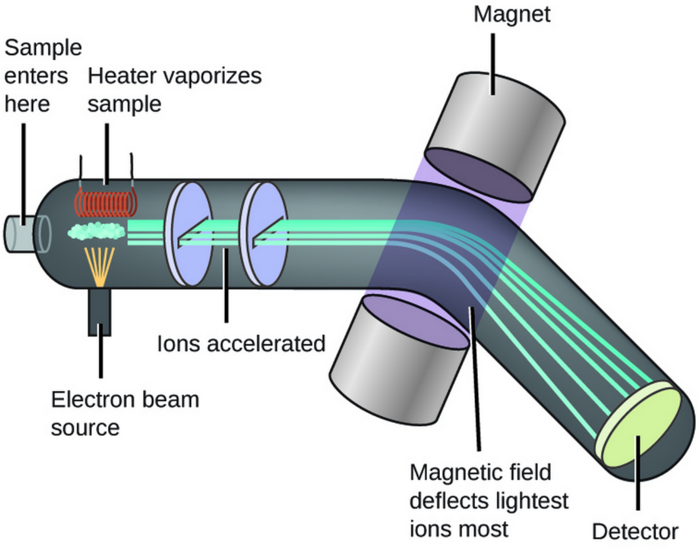


Figure 1.2. A diagram of a mass spectrometer. A sample is injected into the machine, vaporized by a heater, and then ionized by a stream of high-energy electrons. The resulting ions are accelerated through parallel electric plates and then deflected in a magnetic field before they reach a detector. Image credit: "Atomic Structure and Symbolism: Figure 5" by OpenStax Chemistry, CC BY 4.0.

technologies which are central to their ability to analyze and identify compounds in samples with high precision [4].

While this simple setup illustrates the core operations of an MS experiment, it is insufficient for real biological samples due to their complexity. Whole-cell lysates contain tens of thousands of different peptide species, far exceeding the resolving power and dynamic range of current instruments. Therefore, separation techniques are introduced prior to MS analysis. High-performance liquid chromatography (HPLC) provides time-resolved separation on the scale of seconds, effectively reducing sample complexity. Additionally, Ion Mobility Spectrometry (IMS) may be employed to further separate ions based on their shape and charge within milliseconds [5]. These orthogonal separations improve signal clarity and proteome coverage.

Moreover, a single MS spectrum of a peptide ion reveals only its molecular weight, which is often insufficient for confident identification. To extract sequence information, selected peptide ions undergo fragmentation, producing smaller ions that are analyzed in a second stage of mass spectrometry (MS/MS). The resulting fragment spectra can be interpreted computationally to reconstruct the peptide sequence and identify the originating protein [1], [6]. The integration of all these steps—enzymatic digestion, separation, ionization, mass analysis, and fragmentation—constitutes the core of a typical bottom-up proteomics workflow. When this approach is applied to complex protein mixtures such as whole-cell lysates without targeting specific proteins, it is referred to as shotgun proteomics. This strategy allows for the large-scale, unbiased identification and quantification of thousands of proteins in a single experiment and has become a cornerstone of modern proteomics research.

The complexity of biological samples, particularly in proteomics, presents a significant challenge to traditional MS methods. In particular, whole-cell lysates or complex protein mixtures often contain thousands of peptides that overlap in terms of their mass-to-charge (m/z) ratios, making it difficult to resolve individual ions with sufficient accuracy. The following section will explore the principles and applications of an advanced approach for proteomics, which addresses these issues.

1.1.3 An Advanced Workflow for Proteomics: high-performance liquid chromatography, electrospray ionization, ion mobility spectrometry, and tandem mass spectrometry

In its more general setup, shotgun proteomics combines three levels of separation: High-Performance Liquid Chromatography (HPLC), IMS, and MS, with Electrospray Ionization (ESI) acting as the interface between the liquid and gas phases. This integrated workflow, termed HPLC-ESI-IMS-MS/MS, enhances the sensitivity and resolution of proteomics analysis by providing orthogonal separation strategies. As peptides elute from the HPLC column in the liquid phase, they are ionized by ESI, which gently converts them into gas-phase ions suitable for analysis by IMS and MS. Each level of separation contributes to reducing sample complexity and improving the clarity of the resulting spectra, making it easier to identify and quantify peptides in highly complex biological samples.

High-Performance Liquid Chromatography

HPLC is a widely used technique in proteomics due to its ability to separate complex peptide mixtures based on their chemical properties, such as polarity and size [7]. Fig.1.3 shows a representation of a state-of-the-art HPLC system. The stationary phase(gray background) is the solid or liquid phase that is fixed in place inside the chromatographic column, while the mobile phase (red and yellow dots) is the solvent that moves through the column. As the peptide mixture is introduced into the column, the peptides interact with the stationary phase, and their movement through the column is governed by the interactions between the peptides and the mobile and stationary phases. The stationary phase often consists of small particles, typically made from silica or polymer, coated with specific functional groups designed to interact with the peptides [8].

The separation in HPLC occurs over the order of seconds, with peptides eluting at different times depending on their chemical characteristics. HPLC operates based on a variety of physical phenomena, including hydrophobic interactions, electrostatic forces, and hydrogen bonding [9]. Hydrophobic interactions occur between non-polar peptides and the non-polar surface of the stationary phase, slowing their movement through the column. Electrostatic interactions arise when peptides with charged functional groups (such as basic or acidic side chains) interact with charged groups on the stationary phase, affecting their retention time [9]. Hydrogen bonding between polar peptide functional groups and the stationary phase further influences peptide movement through the column [8].

In reversed-phase HPLC, which is the most common mode used in proteomics, the stationary phase is non-polar, often consisting of long hydrocarbon chains (such as C18), while the mobile phase is more polar. The term "reversed-phase" refers to the reversal of polarity between the stationary and mobile phases, compared to

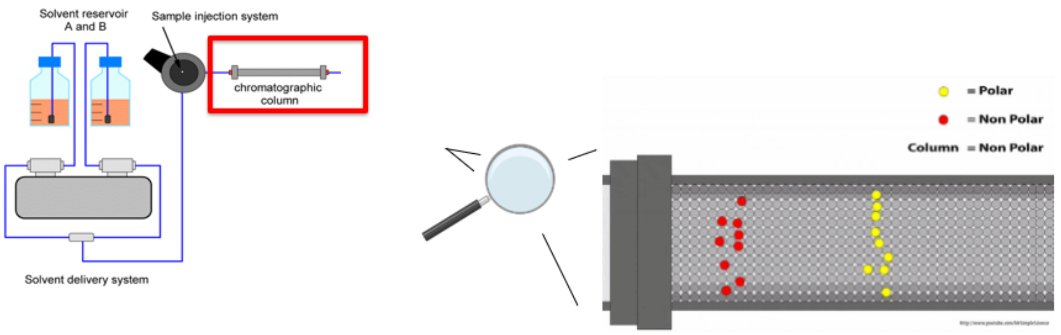


Figure 1.3. Schematic Layout of a Reversed-phase HPLC system. Adapted from: Oliver Scherf-Clavel, "Impurity Profiling of Challenging Active Pharmaceutical Ingredients without Chromophore," Ph.D. Thesis, Ludwig Maximilian University of Munich and from: MrSimple-Science, "HPLC - The Stationary Phase - Animated", YouTube, timestamp 2:06.

normal-phase HPLC, where the stationary phase is polar and the mobile phase is non-polar [7]. In reversed-phase HPLC, hydrophobic peptides (non-polar side chains) interact more strongly with the hydrophobic stationary phase and thus elute later, as they are retained for a longer time before being carried away by the polar mobile phase. In contrast, hydrophilic peptides (more polar) interact less with the stationary phase and pass through the column more quickly, eluting earlier [7]. The efficiency of separation depends on factors such as the nature of the stationary phase, the composition of the mobile phase, the flow rate, and the temperature.

The role of HPLC in the HPLC-ESI-IMS-MS/MS workflow is to reduce sample complexity by narrowing the range of peptides before they enter the ionization stage. By isolating peptides into narrower fractions, HPLC reduces the likelihood of co-elution and ion interference during subsequent stages of analysis [9].

Electrospray Ionization

ESI is a soft ionization technique that plays a central role in interfacing liquid-phase separations with gas-phase mass spectrometry. In shotgun proteomics workflows, ESI is typically employed immediately downstream of HPLC to convert eluting peptides from solution into gas-phase ions suitable for mass analysis. This transition is essential, as both IMS and MS operate on charged, gas-phase analytes.

The core mechanism of ESI relies on the application of a high electric potential (typically 2–5 kV) to the tip of a conductive capillary through which the liquid sample flows. This creates a strong electric field at the solvent–air interface, leading to the formation of a Taylor cone and subsequent ejection of charged droplets into the surrounding atmosphere [10]. These droplets undergo rapid solvent evaporation assisted by a nebulizing gas and thermal desolvation. As the droplets shrink, charge density increases until Coulombic repulsion causes fission into smaller daughter droplets. Through successive evaporation and Coulomb fission cycles—each triggered when droplets exceed the Rayleigh limit—the process gradually produces increasingly charged, smaller droplets, eventually liberating free gas-phase analyte ions that represent the original solutes [11].

Figure 1.4 illustrates the ESI process in positive ion mode. A high voltage is applied

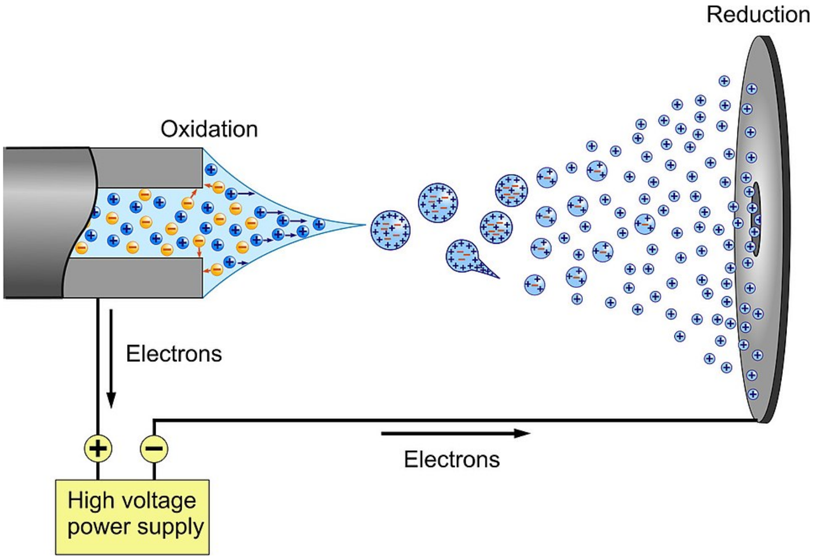


Figure 1.4. Schematic of ESI in positive mode. A high-voltage potential generates a Taylor cone at the emitter tip, forming charged droplets that undergo desolvation and Coulombic fission, ultimately producing gas-phase ions for MS analysis. Image by Lukasz Dziubek, licensed under CC BY 2.0 https://commons.wikimedia.org/wiki/File:ESI_positive_mode_(21589986840).jpg.

to a liquid stream exiting a narrow-bore emitter, causing the formation of a Taylor cone from which a fine spray of charged droplets emerges. These droplets are desolvated through the application of heat and a drying gas, gradually producing smaller droplets and, eventually, free gas-phase ions. These ions are then directed into the mass spectrometer through an orifice held at reduced pressure. The image captures the essential stages of ion formation and transfer, providing a clear visualization of how ESI bridges liquid-phase separation and gas-phase detection.

A key advantage of ESI is its ability to produce multiply charged peptide ions—enabling efficient analysis of large biomolecules on instruments with limited m/z ranges—thanks to its mechanism of transferring multiple protons per peptide [12]. Moreover, ESI's compatibility with volatile aqueous–organic mobile phases used in reversed-phase HPLC ensures seamless, online coupling critical for high-throughput proteomics workflows [13]. Importantly, ESI is considered a "soft" ionization method because it typically does not fragment peptides during ionization, thus preserving their structural integrity for downstream tandem mass spectrometry analysis.

Several ESI configurations exist, including nano-ESI, which operates at lower flow rates (nL/min scale) and offers enhanced sensitivity for limited sample amounts [14]. In the context of bottom-up proteomics, ESI enables continuous, online coupling between HPLC and MS or HPLC-ESI-IMS-MS/MS setups, maintaining temporal resolution and minimizing sample loss.

Overall, electrospray ionization serves as the critical bridge between chromatographic separation and mass-based detection, ensuring that peptides are efficiently transferred, ionized, and preserved for high-resolution proteomic analysis [12].

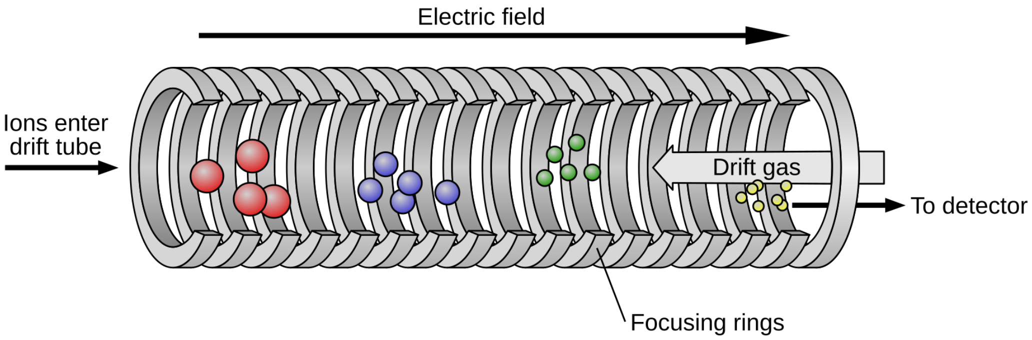


Figure 1.5. Diagram of an ion mobility spectrometry setup illustrating the ion gate, drift region, applied electric field, and buffer gas collisions that separate ions by size, shape, and charge. Image by Jeff Dahl, licensed ucer CC BY-SA3.0 https://en.wikipedia.org/wiki/File:Ion_mobility_spectrometry_diagram.svg

ION MOBILITY SPECTROMETRY

After separation by HPLC and ionization, peptides are introduced into an IMS, where they undergo further separation based on their shape, size, charge state, and Collision Cross Section (CCS). IMS operates in the order of milliseconds, offering a complementary separation method to HPLC that works on a much faster timescale. This makes IMS an excellent choice to couple with HPLC, as it can efficiently resolve ions that elute from the chromatographic column and separate them in a much shorter time, thus providing additional resolution [15].

IMS separates ions based on their mobility in an electric field as they travel through a gas, see Fig.1.5. The ion drift time is determined by the size, shape, and charge of the ion. Smaller, more compact ions experience less resistance from the gas molecules and thus travel faster, while larger or more irregularly shaped ions experience more drag, leading to slower drift times. The drift time can be measured with high precision and serves as a characteristic feature of each ion. Furthermore, ions may have more than one peak in the IMS spectrum, corresponding to different conformations or charge states [16]. This makes IMS particularly useful for analyzing peptide ions that may exist in multiple charge states, providing additional structural information that is not available from MS alone.

The fundamental parameter measured by IMS is the mobility of ions, which is the rate at which ions move through the drift tube in response to an applied electric field. Ion mobility is influenced by the ion's charge, shape, and size, and is often expressed in terms of the CCS (Ω), which describes the effective area over which the ion interacts with the gas molecules. Ions with larger CCS will experience more resistance, resulting in slower drift times. This allows IMS to separate ions not just by their mass-to-charge ratio (m/z) but by their 3D shape and size as well. The collision cross-section is a key structural parameter, offering insight into the conformation of peptides or proteins that may otherwise be indistinguishable in traditional mass spectrometry [17]. There are studies that offer a compendium of a vast array of CCS values obtained through IMS [18].

In IMS, the low-field regime refers to the situation where ions are subjected to weak electric fields, typically at voltages less than 10 V/cm. In this regime, the ions move through the gas at a rate where the applied electric field is not strong enough to cause significant ion-molecule collisions that would affect the ion's trajectory in the drift tube. In the low-field regime, ion mobility is primarily determined by the size, shape, and charge of the ions, and the behavior of the ions follows a linear relationship with respect to the applied electric field. The Mason–Schamp equation is often used to describe the ion mobility in this regime. It is given by:

$$K = \frac{Z}{\eta} \cdot \left(\frac{2e}{3k_B T}\right)^{1/2} \cdot \left(\frac{1}{m^{1/2}}\right) \tag{1.6}$$

where K is the ion mobility, Z is the ion's charge, η is the gas viscosity, e is the electron charge, k_B is Boltzmann's constant, T is the temperature, and m is the ion's mass. The equation describes how ion mobility depends on the size and charge of the ion as well as the properties of the gas [19]. This relationship is crucial in determining how ions drift through the gas and is particularly valuable for separating ions that are otherwise indistinguishable in traditional mass spectrometry.

IMS is particularly useful when analyzing isomeric peptides or peptides with similar masses, as it provides a third dimension of separation beyond mass and charge, improving the overall resolution of the analysis. By separating ions based on their size, shape, and charge, IMS enhances the ability to identify and quantify peptides, especially in complex mixtures where peptides might otherwise co-elute in HPLC or have overlapping m/z values in MS.

Mass Spectrometry

MS is a key component in the HPLC-ESI-IMS-MS/MS workflow, where it plays an essential role in identifying and quantifying peptides based on their mass-to-charge (m/z) ratios. MS is a combination of two primary components: the mass analyzer and the detector. The mass analyzer is responsible for separating ions based on their m/z ratios, while the detector measures the abundance of each ion species. In the context of the HPLC-ESI-IMS-MS/MS workflow, MS serves as the initial survey stage, generating a broad overview of the peptide ions present in the sample.

In this step, the mass spectrometer creates a 4D peak in the Retention Time (RT), IMS, m/z and intensity space. This multidimensional dataset can be visualized in parts using 3D projections. Figure 1.6 illustrates two common representations. Panel (a) shows a 3D projection of LC-MS data, where intensity is plotted against retention time and m/z, creating a series of chromatographic peaks resolved by mass analysis. Panel (b) extends this concept by adding ion mobility as a third separation axis, forming a 4D landscape that improves ion resolution and separation in complex samples.

The resolution of the mass spectrometer plays a crucial role in how well it can distinguish between ions with very similar m/z values. Higher-resolution mass spectrometers can more precisely differentiate ions with small differences in m/z, which is essential for identifying peptides in complex mixtures. Mass spectrometers typically offer resolutions ranging from a few thousand to several hundred thousand, depending on the design of the instrument. The higher the resolution, the more accurately the ions can be distinguished from each other, and the more detailed the

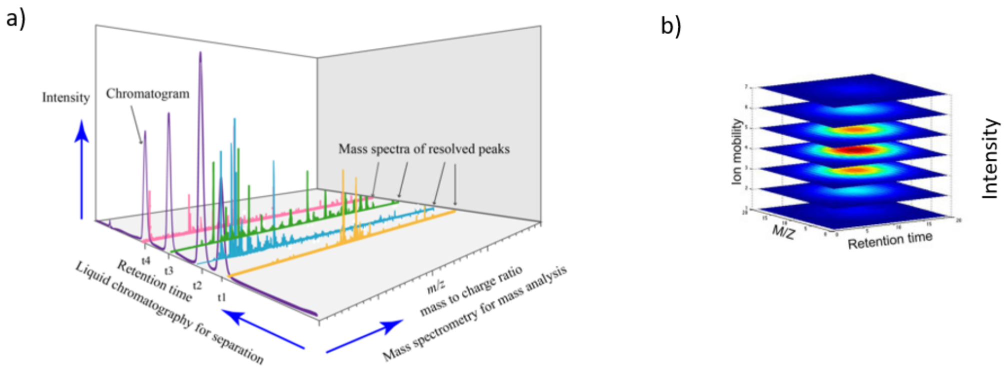


Figure 1.6. Visualization of MS signal in LC-MS(/IMS) workflows. (a) Intensity as a function of retention time and m/z, showing chromatographic separation and resulting mass spectra of resolved peaks. (b) A 3D representation adding ion mobility as a separation dimension to distinguish co-eluting and isobaric ions. Adapted from Wikimedia Commons, CC BY-SA 4.0 https://en.wikipedia.org/wiki/File:Liquid_chromatography_MS_spectrum_3D_analysis.png and from *Molecular and Cellular Proteomics* (doi:10.1074/mcp.TIR119.001720.)

spectra will be. High-resolution instruments, such as Orbitraps or Fourier Transform Ion Cyclotron Resonance, provide extremely accurate m/z measurements, which is particularly important when analyzing complex proteomic data sets.

The timescale at which MS operates is crucial for understanding how quickly data is acquired. In an HPLC-ESI-IMS-MS/MS workflow, HPLC typically operates on the scale of seconds, separating peptides before they enter the IMS and mass spectrometer. IMS happens on the order of milliseconds, providing a rapid additional separation step. The MS survey itself can also take place within milliseconds to seconds, depending on the instrument's capabilities and the complexity of the sample. The entire process from HPLC separation to MS analysis is incredibly fast, enabling high-throughput analysis of complex proteomes.

While this MS survey provides valuable information, the resulting spectra are still quite convoluted due to the complexity of the sample. Multiple ions may have overlapping m/z values, and the chromatographic peaks can also overlap, making it difficult to distinguish individual peptides. Despite this, the mass spectrometer is capable of identifying peaks that correspond to unique peptide species, even if they are not fully resolved. Based on this initial survey, the most abundant and best-resolved ions are selected for further sequencing in the MS/MS stage [20].

Tandem Mass Spectrometry (MS/MS)

The MS/MS stage is the critical stage where the detailed sequencing of peptides occurs and is performed after the initial MS analysis. This process is illustrated in Figure 1.7, which shows how precursor ions from the MS1 scan are selected for fragmentation and further analysis in MS/MS. In the MS1 3D spectrum on the left, peaks represent different ion species across retention time and m/z. The colored squares (red, blue, and green) indicate narrow isolation windows—typically around 0.5 Da wide—used to target specific precursor ions for fragmentation. These windows ensure precise selection of individual ions for MS/MS analysis, minimizing interference from neighboring species. This selection step is critical for obtaining high-quality sequence information in the subsequent fragmentation stage. The Dalton (Da) is the unit of mass used in MS, defined as one twelfth the mass of a carbon-12 atom, or approximately 1.66×10^{-24} grams.

The MS/MS spectra produced in this stage are one-dimensional and represent the fragmentation pattern of the selected peptide ion. These spectra correspond to the right-hand side of Figure 1.7, where each panel illustrates the fragment ion spectrum resulting from the fragmentation of a single precursor ion selected in MS1. The intensities in each MS/MS spectrum represent the probability that a specific bond within the peptide molecule will break during the fragmentation process. The stronger the intensity of a fragment peak, the more likely it is that the corresponding bond broke during ionization. Additionally, losses (such as water or ammonia) and contaminants (such as background noise from the sample matrix or the instrument) may appear in the spectra, and these need to be interpreted carefully by software or an experienced researcher.

This fragmentation process is essential because it enables the sequencing of peptides. By breaking the peptide into smaller fragments and analyzing their m/z ratios, the mass spectrometer can "read" the sequence of the original peptide. Specific bonds between amino acids (peptide bonds) are more likely to break under Collision-Induced Dissociation (CID), and this process follows predictable patterns. Therefore, the fragment ion patterns provide consistent clues about the peptide's sequence.

The fragmentation in MS/MS is not random, it tends to occur at predictable sites, such as peptide bonds, especially under CID,, making the resulting ion patterns consistent and highly informative for peptide sequencing. The peptide is typically broken into y-ions (from the C-terminal) and b-ions (from the N-terminal), with each fragment corresponding to a part of the peptide sequence. These specific fragments are targeted because they produce clear, structured patterns that can be interpreted to deduce the peptide sequence. For example, the b-ions and y-ions provide insight into how the peptide breaks when specific bonds between amino acids are cleaved: the y-ion corresponds to the fragment containing the C-terminal part of the peptide, while the b-ion contains the N-terminal portion. By breaking the peptide into these fragments and analyzing their m/z ratios, the mass spectrometer can effectively "read" the sequence of the original peptide.

The MS/MS stage differs significantly from the initial MS1 survey stage in terms of both instrumentation and speed. During MS1, the mass spectrometer conducts a broad survey to collect all the ions in the sample, creating a 3D peak in the RT, IMS, and m/z space. In contrast, MS/MS focuses on analyzing a single peptide species selected from the MS1 scan, performing ion fragmentation and analyzing the resulting fragments.

This requires different instrumentation, particularly in how the ions are isolated and fragmented. The MS1 stage typically uses a mass analyzer like a quadrupole, which scans over a broad range of m/z values to detect all ions. In MS/MS, however, a more specialized system is employed that can select a single ion, fragment it, and then analyze the resulting fragments. The most commonly used technique for this

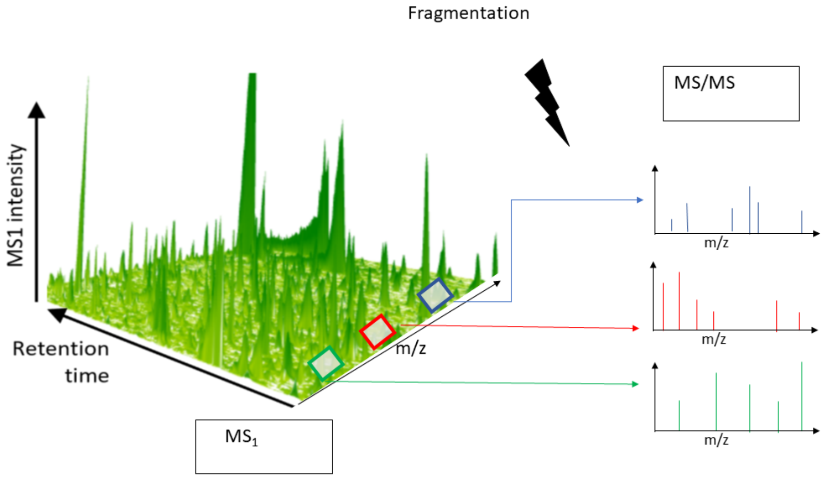


Figure 1.7. Illustration of tandem mass spectrometry workflow. The MS1 survey scan shows a 3D plot of retention time, m/z, and intensity. Selected precursor ions (highlighted in red, blue, and green boxes) are isolated and fragmented, resulting in MS/MS spectra used for peptide sequencing. Figure adapted from *Nature Biotechnology* (doi:10.1038/s41587-022-01424-w).

is CID,. This step typically occurs in a trap cell, where the selected ion is trapped and then subjected to collisions with an inert gas (usually nitrogen or argon). These collisions cause the peptide to break apart into smaller ions, which are then detected in the second stage of the mass spectrometer. The trap cell is specifically designed for this process, whereas the quadrupole in MS1 serves to simply filter ions based on their m/z.

Additionally, the speed of the MS/MS analysis differs from the initial MS stage. MS1 is much faster, as it conducts a broad survey across a range of ions, while MS/MS is slower because it involves more complex processes, including ion isolation, fragmentation, and the collection of fragment spectra. The MS/MS analyzer must be able to isolate specific ions and perform these fragmentation processes efficiently, and this typically involves slower scanning speeds compared to the MS1 survey mode.

In summary, the MS/MS analyzer needs to be more specialized and precise than the MS1 analyzer. It must be capable of isolating a single peptide ion, fragmenting it with high energy, and analyzing the resulting fragments with high resolution. This process is slower but provides the detailed peptide sequence data necessary for identifying proteins in complex samples.

1.2 Acquisition Strategies

The strategy used to select peptide ions for fragmentation in MS/MS has a significant impact on the quality and reproducibility of proteomics data. The main acquisition methods—Data-Dependent Acquisition (DDA), Data-Independent Acquisition (DIA), and Targeted Proteomics—differ in how precursor ions are selected and how fragmentation data is collected. Figure 1.8 visually compares these acquisition strategies.

Each panel displays a heatmap of precursor ions over time (y-axis) and m/z (x-axis), overlaid with pink rectangles representing MS/MS acquisition events. In the left panel (

Data-Dependent Acquisition has historically been the most widely used method in proteomics. In DDA, the mass spectrometer performs a survey scan to detect all ions and then selects the most abundant precursors for fragmentation in real time. This approach is relatively simple to implement and integrates well with database search tools, which contributed to its widespread adoption. However, DDA suffers from stochastic sampling, meaning that low-abundance peptides are often missed, and identifications may vary between replicates [21]. Despite this, DDA remains a powerful method for exploratory proteomics and has been instrumental in early large-scale proteome studies.

Data-Independent Acquisition emerged as a solution to the limitations of DDA. In DIA, the instrument fragments all precursor ions within a series of predefined m/z windows, ensuring systematic and comprehensive sampling of the entire sample. This acquisition mode enhances reproducibility and quantification accuracy, especially for low-abundance peptides [22]. DIA requires more sophisticated computational tools to deconvolute the highly multiplexed spectra, but recent advances have made this tractable. Techniques such as SWATH-MS have become emblematic of DIA's growing dominance, offering the ability to consistently quantify thousands of peptides across samples with minimal missing data [23]. DIA is increasingly preferred for large cohort studies and biomarker validation.

Targeted Proteomics, including methods like Selected Reaction Monitoring and Parallel Reaction Monitoring, focuses on pre-selected peptides of interest. These methods offer high sensitivity, specificity, and quantitative precision, making them ideal for validating biomarkers, studying specific pathways, and supporting clinical assays [24]. Although targeted approaches do not offer proteome-wide coverage, they provide robust measurements where reproducibility is essential.

The historical evolution from DDA to DIA and targeted methods reflects the increasing demand for sensitivity, reproducibility, and quantifiability in proteomics. The choice of acquisition strategy must align with the study design—whether exploratory, confirmatory, or clinical—and often involves balancing depth of coverage against data complexity and analysis burden.

Even with advanced acquisition strategies and upstream deconvolution techniques like chromatographic and ion mobility separation, the output of a proteomics experiment remains highly complex. The raw data consists of a large number of precursor and fragment ion spectra that must be interpreted and organized. To transform this data into meaningful biological insight, sophisticated data analysis tools are required. These tools aim to extract a high-confidence list of identified and quantified proteins by performing a series of computational steps including spectral preprocessing, peptide-spectrum matching through database searches, and rigorous False Discovery Rate (FDR) control. The next section will explore the main stages of data analysis in proteomics, explaining the algorithms and methodologies used to translate raw MS data into robust biological conclusions.

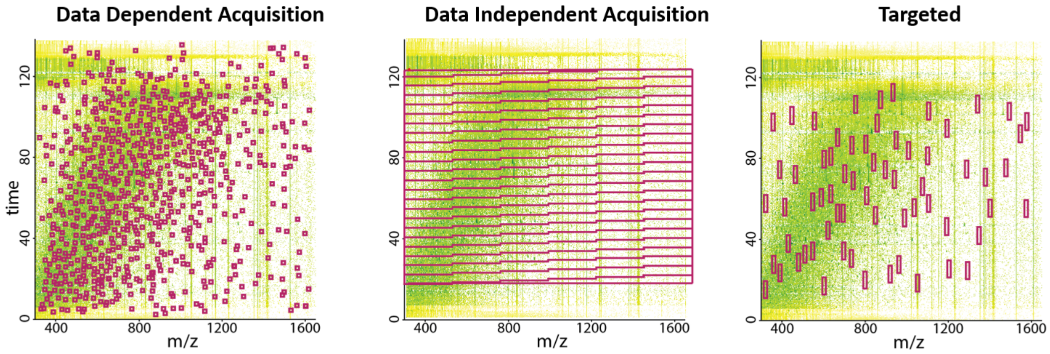


Figure 1.8. Comparison of acquisition strategies in LC-MS/MS workflows. Each panel shows the selection of precursor ions (pink boxes) in the time versus m/z space. Left: Data-Dependent Acquisition selects precursors based on abundance. Center: Data-Independent Acquisition scans predefined m/z windows across time. Right: Targeted acquisition focuses on preselected m/z-retention time regions. Figure adapted from *Annual Review of Biomedical Data Science* (doi:10.1146/annurev-biodatasci-080917-013516).

1.3 Data Analysis

The ultimate goal of data analysis in proteomics is to obtain a high-confidence list of identified and quantified proteins from complex raw mass spectrometry data. This process is critical for interpreting biological phenomena and drawing meaningful conclusions from proteomics experiments. Due to the complexity of the data—even after upstream simplifications such as chromatographic separation and advanced acquisition strategies—robust computational workflows are required to process, interpret, and validate the results.

Data analysis typically follows three main stages: preprocessing, peptide-spectrum matching, and postprocessing. Figure 1.9 provides an overview of these stages, from initial spectral data refinement to identification and quantification.

The left panel of the figure illustrates typical preprocessing operations applied to raw spectral data. These include centroiding, which converts profile spectra into simplified peak representations; deisotoping, which removes redundant isotopic peaks to retain only monoisotopic signals; and collapsing charge states to unify multiply charged signals of the same peptide into a single mass representation. These transformations simplify the data structure while preserving key features used for matching and quantification.

The center panel of Figure 1.9 illustrates the core step of Peptide-Spectrum Match (PSM), where experimental MS/MS spectra are interpreted by assigning them to candidate peptide sequences. This step is critical because it directly determines which peptides—and consequently, which proteins—are identified in the sample. Accurate PSM ensures meaningful biological interpretations and high-confidence identifications.

The approach to PSM differs significantly between DDA and DIA workflows. In DDA, each MS/MS spectrum corresponds to a relatively clean fragmentation event of a single precursor ion, allowing the use of sequence database search (center panel of Figure 1.9, right branch) . In this approach, the software generates theoretical spectra from in silico–digested peptides and compares them to the observed spectra to find the best match. One of the most widely used software tools for DDA analysis is MaxQuant, which integrates all steps from raw data handling to protein quantification and statistical analysis, and has been foundational in large-scale quantitative proteomics studies [25]. MaxQuant uses the Andromeda search engine [26] for peptide identification and includes advanced algorithms for controlling the statistical significance of the identifications at both the peptide and protein levels. Other prominent tools include FragPipe, which incorporates MSFragger for ultrafast database searching and sophisticated quantification strategies, and has gained popularity for its speed and open-source accessibility [27], and Proteome Discoverer by Thermo Scientific, a modular platform supporting diverse workflows and search engines[28].

In contrast, DIA produces highly multiplexed MS/MS spectra containing fragments from multiple precursors simultaneously. As a result, DIA relies on spectral library search(center panel of Figure 1.9, left branch), which matches observed spectra against a curated collection of empirically acquired peptide spectra. However, acquiring high-quality empirical spectral libraries can be resource-intensive and not always feasible for every experimental condition. To address this, state-of-the-art DIA tools such as DIA-NN[29], Spectronaut [30], and MaxDIA [31] incorporate functionality to predict spectral libraries in silico. These tools leverage peptide fragmentation models and machine learning to simulate theoretical spectra directly from sequence databases, reducing the dependency on empirical libraries and improving flexibility and accessibility in DIA analysis.

MaxDIA works by extending the MaxQuant platform for DIA data, integrating the familiar MaxQuant interface and statistical framework with a library-free, direct DIA approach. It uses a predictive model to generate in silico spectral libraries based on the sample's FASTA file and peptide fragmentation rules, bypassing the need for pre-constructed empirical libraries. MaxDIA aligns the predicted library to the measured data by bootstrapping identifications to fit recalibration functions of increasing complexity until they become comparable. Once aligned, candidate PSMs are detected within specified mass and retention time tolerance windows. These candidate matches are then rescored by incorporating various metadata features, including retention time prediction, precursor intensity, and co-elution profiles. A machine learning-based classification step is used to distinguish true from false identifications by training a model on both target and artificially generated decoys. This is done using cross-validation to prevent overfitting and to maintain generalization. FDR control is applied on each cross-validation split to the ranked list to select a set of high-confidence identifications suitable for downstream quantification and biological interpretation. This streamlined process allows users to analyze DIA data within the same environment as DDA, facilitating consistent data processing and interpretation across acquisition methods.

The right panel of Figure 1.9 summarizes essential postprocessing tasks. FDR control is commonly implemented using a decoy database strategy, where reversed or scrambled versions of the original sequences are included in the search to estimate the rate of incorrect identifications thus retaining only high-confidence identifications. Quantification, shown below, aggregates peptide intensities across multiple samples to infer relative or absolute protein abundance. Together, these steps ensure that final results are both statistically rigorous and biologically meaningful.

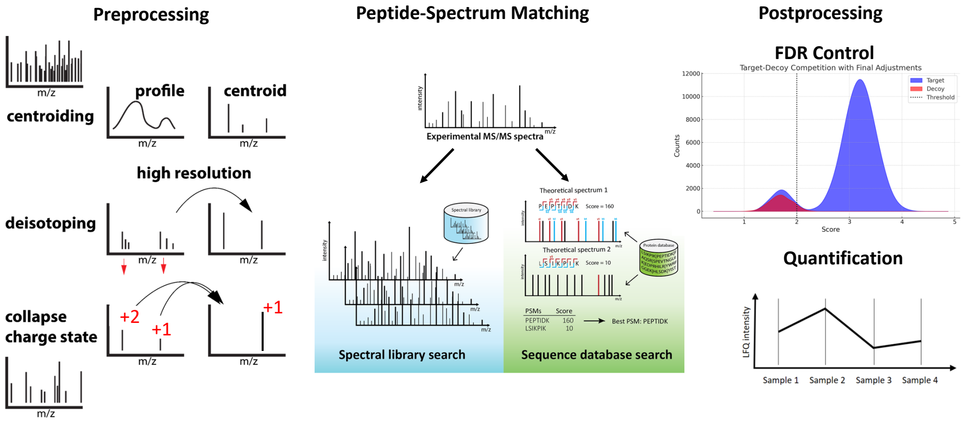


Figure 1.9. Overview of a typical data analysis workflow in proteomics. Left: Preprocessing transforms raw spectral data into simplified and unified representations suitable for identification. Center: Peptide-spectrum matching through spectral library search and sequence database search assigns peptides to experimental MS/MS spectra. Right: Postprocessing applies statistical validation (e.g., FDR control) and protein quantification. Middle panel adapted from Gutenbrunner, P. (2022), doctoral thesis, LMU Munich https://edoc.ub.uni-muenchen.de/29625/1/Gutenbrunner_Petra.pdf

2
Deep learning in proteomics

2.1 Introduction to Deep Learning

Deep Learning (DL) is a subfield of Machine Learning (ML) focused on algorithms inspired by the structure and function of the brain, known as artificial neural networks. While the concept of artifical neural networks dates back to the mid-20th century, it is only in the past decade that advances in computational power, algorithmic innovation, and data availability have enabled deep learning to become a transformative tool across many scientific disciplines [32].

At its core, DL involves training layered architectures—typically composed of multiple interconnected neurons or nodes—to automatically learn representations from data. These models are particularly well-suited to handling complex, high-dimensional inputs, and they excel at uncovering hierarchical patterns that are difficult to detect with traditional statistical methods. DL has revolutionized fields such as computer vision, natural language processing, and speech recognition [33][34], and it is increasingly finding application in the life sciences, including genomics, structural biology, and proteomics[35].

A key strength of DL is its ability to perform end-to-end learning: raw inputs can be fed into the model, which then learns both the features and decision rules simultaneously. This stands in contrast to classical ML pipelines, which typically require hand-crafted feature engineering.

In proteomics, the growing complexity and volume of data—ranging from raw mass spectra to protein sequences and structural annotations—make deep learning a natural candidate for advancing computational methods. Applications of deep learning in this field include

- predicting peptide fragmentation patterns[36], RT[37], and post-translational modifications
- enhancing protein identification and quantification
- and developing generative models for peptide design and de novo sequencing [38][39]

The following sections will explore two major architectures—Recurrent Neural Network (RNN) and transformers—and how they are being applied to tackle key challenges in proteomics.

2.2 Recurrent Neural Networks

RNN are a class of neural networks specifically designed to handle sequential data. Unlike traditional feedforward networks, RNNs possess an internal state (or memory) that allows them to retain information from previous inputs, making them suitable for tasks where the order and context of inputs matter. They have the ability to model temporal dependencies by iterating over input sequences one element at a time while maintaining a hidden state. This structure naturally encodes positional information and context, allowing RNNs to capture both local and long-range dependencies in data[40]. This makes RNNs particularly effective for processing data of variable length, such as text, time series, or biological sequences.

In proteomics, RNNs are widely used as they offer a natural way for processing peptide sequences. They have been employed in tasks such as de novo peptide sequencing, modeling peptide fragmentation patterns, predicting retention times, and generating synthetic peptides. Their ability to understand amino acid sequences as temporal data allows for nuanced modeling of biochemical properties and mass spectrometric behavior [38][39][37][41][42].

RNNs are widely supported by modern deep learning frameworks such as TensorFlow [43] and PyTorch [44], which provide built-in modules for specific RNNs architectures. These frameworks enable efficient model development, training, and deployment, making them highly accessible for researchers in the life sciences.

Several variants of the basic RNN cell have been developed to address issues such as vanishing or exploding gradients during training. The most prominent among these are Long Short-Term Memory (LSTM) cells[45] and Gated Recurrent Units (GRU)s[46]. These architectures introduce gating mechanisms that help retain relevant information over long sequences and suppress irrelevant signals.

2.2.1 LSTM Architecture

Figure 2.1 illustrates the internal structure of a LSTM unit, which was designed to overcome the vanishing gradient problem common in traditional RNNs. Unlike a standard RNN cell, the LSTM cell includes three primary gates—**the forget gate**, **input gate**, and **output gate**—which control the flow of information and enable the network to maintain long-term dependencies

As shown in the diagram, each LSTM unit operates on the input x_t, the previous hidden state h_{t-1}, and the previous cell state c_{t-1}. These gates manage the internal cell state c_t and the hidden state h_t as follows:

- **Forget Gate** (leftmost orange box in the diagram): Computes $f_t = \sigma(W_f \cdot [h_{t-1}, x_t])$, where σ is the sigmoid function. This gate determines which parts of the previous cell state c_{t-1} should be forgotten (element-wise multiplication).

- **Input Gate** (middle orange box): Includes two components—$i_t = \sigma(W_i \cdot [h_{t-1}, x_t])$ and $\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t])$. The sigmoid gate i_t controls how much new information to add, and the candidate cell state $\tilde{c}_t$ provides the new values to be potentially added. The updated cell state is computed as $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$, where $\odot$ denotes element-wise multiplication.

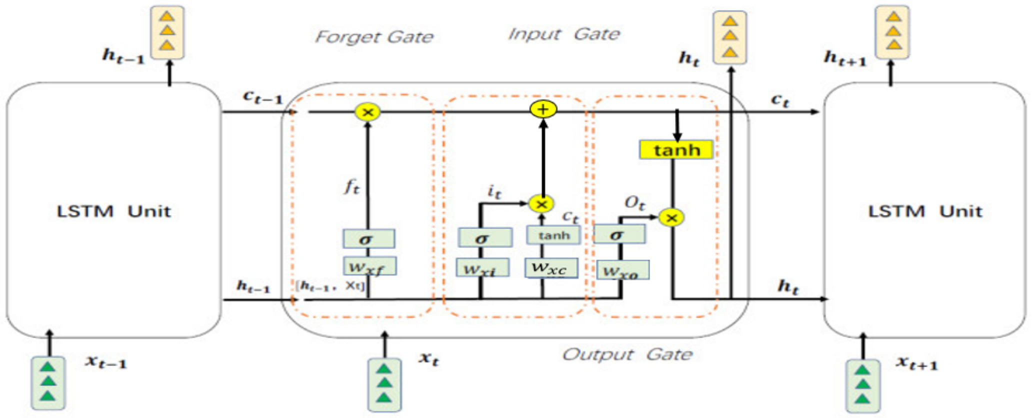


Figure 2.1. Internal architecture of an LSTM unit showing the flow of information through the forget, input, and output gates. Each gate controls updates to the cell state c_t and the hidden state h_t using combinations of the previous hidden state h_{t-1} and the current input x_t. Image adapted from Liu et al., *Mathematical Biosciences and Engineering*, 20(9):17569-17588, 2023 http://dx.doi.org/10.3934/mbe.2023780

- **Output Gate** (rightmost orange box): Determines the new hidden state h_t using $o_t = \sigma(W_o \cdot [h_{t-1}, x_t])$, followed by applying $\tanh(c_t)$ to the new cell state. The hidden state is then $h_t = o_t \odot \tanh(c_t)$, controlling which parts of the cell state are revealed to the next layer or time step.

This gated architecture allows the LSTM to regulate the information flow efficiently, preserving relevant information across long sequences and discarding irrelevant parts. The explicit handling of memory through c_t and gating mechanisms makes LSTMs highly suitable for sequence modeling tasks in proteomics, such as learning patterns across peptide sequences or interpreting temporal dependencies in mass spectrometry runs.

To fully leverage the power of LSTMs, the following section will delve into training techniques—covering loss functions, optimization algorithms, and regularization strategies tailored for high-dimensional, sparse biological data.

2.2.2 Training LSTMs

Training LSTM networks involves optimizing the model's weights to minimize a loss function that quantifies the discrepancy between predicted and actual outputs. This process relies on backpropagation through time (BPTT), an extension of the standard backpropagation algorithm adapted for recurrent architectures[47].

Figure 2.2 illustrates the algorithm BPTT applied to an LSTM network. The diagram presents the temporal unrolling of the recurrent architecture as a feedback loop, where the same LSTM unit is conceptually replicated across time steps. During the forward pass, the network processes each input sequentially, producing intermediate hidden states and accumulating the total loss. In the backward pass, gradients are computed at each individual time step and propagated in reverse chronological order through

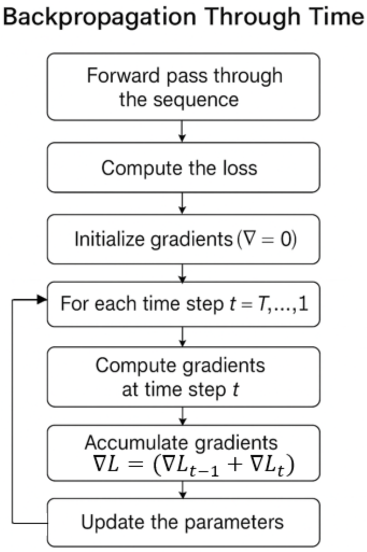

Backpropagation Through Time
Forward pass through the sequence
Compute the loss
Initialize gradients (∇ = 0)
For each time step $t = T,...,1$
Compute gradients at time step t
Accumulate gradients $\nabla L = (\nabla L_{t-1} + \nabla L_t)$
Update the parameters

Figure 2.2. Illustration of the BPTT algorithm applied to LSTM networks. The LSTM is unrolled across several time steps, showing how the loss at each time step contributes to the gradients that are propagated backward through the temporal sequence. This enables the recurrent and input weights to be updated based on their influence across the entire sequence.

the feedback loop. Each time step is treated as a layer in a feedforward network, allowing the application of the chain rule to compute partial derivatives with respect to each parameter. These local gradients are successively accumulated, forming the total gradient used to update the model parameters. This looped structure highlights how BPTT enables credit(or blame) assignment over time, thereby capturing long-term dependencies in sequential data.

The choice of loss function depends on the task. For regression problems such as retention time prediction, the mean squared error is often used. For classification tasks like peptide identification, categorical cross-entropy or negative log-likelihood is more appropriate. In the context of sequence generation, as in de novo peptide sequencing, sequence-level losses and teacher forcing strategies are frequently employed [38][48].

Optimization is typically performed using gradient-based methods such as Adam or RMSprop, which are well-suited for non-stationary objectives and sparse gradients. Learning rate schedules or adaptive learning rate techniques (e.g., AdamW) can help accelerate convergence and avoid local minima. Weight initialization, gradient clipping, and normalization layers are also important for maintaining stable training dynamics.

Overfitting is a common concern when training deep learning models, particularly in proteomics where labeled data may be limited. To mitigate this, regularization techniques such as dropout (applied to inputs and recurrent connections), L2 weight penalties, early stopping, and data augmentation are commonly employed. Batch normalization or layer normalization can further stabilize training, especially in deep or stacked LSTM architectures [49].

Finally, the quality of training data is crucial. Large, diverse peptide datasets with high-confidence annotations from public repositories (e.g., PRIDE or ProteomeXchange) often serve as a foundation for training. Synthetic peptide libraries are also valuable for tasks such as retention time prediction or fragmentation modeling, where high-quality ground truth is essential.

The next section introduces Transformer architectures, which build upon the limitations of RNNs by leveraging self-attention mechanisms to model long-range dependencies more efficiently and in parallel.

2.3 Transformer Models

Transformers represent a major breakthrough in deep learning, particularly for sequence modeling tasks. Introduced by Vaswani et al. in their seminal 2017 paper "Attention is All You Need" [50], transformers fundamentally changed the way models process sequential data. Unlike RNNs, which process inputs sequentially, transformers operate on entire sequences in parallel using a mechanism called self-attention. This allows them to capture both local and global dependencies without the limitations of recurrence.

The self-attention mechanism enables each position in the input to attend to every other position, effectively learning context-aware representations at each layer. Positional encodings are added to the inputs to retain information about token order, compensating for the model's lack of inherent sequential structure. This architecture not only resolves the bottlenecks associated with sequential processing in RNNs but also allows for better scalability, significantly reduced training times, and improved performance on long-range dependency tasks.

Transformers are at the heart of the current deep learning revolution. They power large language models like BERT, GPT, and ChatGPT, which have set new performance benchmarks across a wide range of natural language processing tasks [51][52]. The same architectural principles have been extended beyond text to fields such as vision (e.g., Vision Transformers), biology (e.g., AlphaFold2), and increasingly, proteomics.

In proteomics, transformers are being leveraged for tasks that benefit from their ability to model complex, context-dependent relationships across sequences or spectral data. Applications include peptide and protein sequence modeling, de novo sequencing, fragmentation pattern prediction, and integration of multi-omics data. Their capacity for transfer learning and pretraining on large corpora of biological data makes them especially attractive for data-limited domains like proteomics.

Transformers are considered cutting-edge technology because they combine flexibility, scalability, and generalization in a single unified architecture. Their ability to model global context without relying on recurrence enables them to outperform previous models on a wide range of tasks, while their parallel processing capabilities align well with modern hardware accelerators.

The next sections will dive deeper into the architecture and training of transformer models, followed by their specific applications to proteomics.

2.3.1 Transformer Architecture

At the core of a transformer model is the self-attention mechanism, which allows the model to weigh the relevance of different elements in a sequence when encoding a particular token. This mechanism is implemented via scaled dot-product attention, where each input token is transformed into three vectors: a query (Q), a key (K), and a value (V). The attention weights are computed by taking the dot product of the query with all keys, scaling by the square root of the dimension, and applying a softmax function. The resulting weights are used to combine the value vectors into a context vector that captures the relationships among all tokens.

Figure 2.3 visualizes this in the central panel, where the scaled dot-product attention mechanism is decomposed step-by-step. Inputs Q, K, and V are multiplied and scaled before being passed through a softmax, resulting in weighted values that represent contextualized token representations.

On the right, the figure shows the multi-head attention module, which performs several attention operations in parallel using separate learned projections of Q, K, and V. These parallel outputs are concatenated and linearly transformed to produce the final output. This setup enables the model to capture diverse aspects of token interactions simultaneously.

Transformers consist of stacked layers, each comprising a multi-head self-attention block and a feed-forward neural network. Each layer is followed by layer normalization and residual connections to facilitate training stability and gradient flow, as depicted in the left panel of the figure. In encoder-decoder setups, such as those used for machine translation, the decoder additionally includes masked attention to prevent information leakage from future tokens.

Positional encodings are added to the input embeddings to provide the model with information about the order of the sequence elements, which is otherwise lost due to the model's fully parallel nature. These encodings can be sinusoidal functions or learned embeddings.

In encoder-decoder transformers (such as those used in translation tasks), the encoder processes the input sequence to produce a sequence of context-rich embeddings, which are then consumed by the decoder to generate outputs. In contrast, encoder-only (e.g., BERT) and decoder-only (e.g., GPT) variants simplify the architecture depending on the task.

Transformers' modular, attention-based design allows them to scale efficiently with data and model size, enabling state-of-the-art results in a wide range of domains.

Training well a transformer can be an art. This involves not only selecting appropriate loss functions and optimizers but also managing the considerable computational demands and tuning a large number of hyperparameters. The following section will cover the strategies and considerations involved in training transformers, especially in the context of proteomics applications, where data may be noisy, sparse, or limited in size.

2.3.2 Training Transformers

Training transformers involves several interrelated components that require careful tuning to achieve optimal performance. At the heart of training is minimizing a task-specific loss function using gradient-based optimization. For sequence modeling

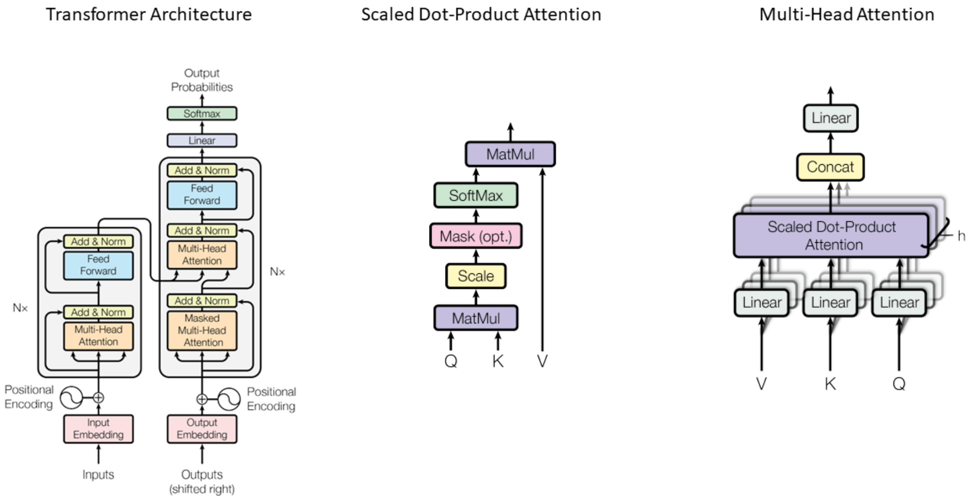


Figure 2.3. Illustration of the transformer model architecture. Left: the encoder and decoder blocks with stacked layers of multi-head attention and feed-forward networks, each followed by residual connections and normalization. Center: the scaled dot-product attention mechanism, which computes attention scores using queries (Q), keys (K), and values (V). Right: the multi-head attention mechanism, which applies multiple parallel attention operations before concatenation and linear projection. Image adapted from Vaswani et al., *Attention is All You Need*, 2017 `https://arxiv.org/abs/1706.03762`.

tasks common in proteomics—such as peptide classification, retention time prediction, or de novo sequencing—cross-entropy and mean squared error are widely used loss functions depending on the output type.

For the gradient-based optimization, the standard backpropagation algorithm is used [53]. During training, the input peptide sequences are tokenized and embedded, augmented with positional encodings, and passed through multiple layers of self-attention and feed-forward sublayers. Gradients are computed for all trainable parameters—including projection matrices for attention (W^Q, W^K, W^V), feed-forward weights, and layer normalization parameters—using automatic differentiation. Optimizers like Adam or AdamW [54] are typically used, often with learning rate warm-up and decay schedules to stabilize training. Attention masking is applied when handling variable-length sequences or enforcing causality in autoregressive tasks. A critical innovation in transformer training is the use of learning rate scheduling strategies—especially the warm-up and decay approach introduced in the original transformer paper [50], which helps stabilize training in early epochs.

Given their depth and number of parameters, transformers are prone to overfitting and require regularization techniques such as dropout[55], label smoothing, and stochastic depth. Additionally, gradient clipping is often employed to prevent exploding gradients.

Transformers also benefit significantly from large-scale pretraining followed by fine-tuning on domain-specific data. In proteomics, this can involve pretraining on massive protein sequence databases and fine-tuning on task-specific datasets such as fragmentation spectra.

A challenge in proteomics is the limited availability of labeled data. To address this, data augmentation strategies—such as in silico spectral simulation, masking of amino acid residues, or shuffling of peptides—can help increase effective training set size and improve generalization.

Training transformers also places high computational demands. Distributed training, mixed-precision arithmetic[56], and memory-efficient implementations (e.g., FlashAttention[57], DeepSpeed[58]) are commonly used to accelerate model convergence while fitting into hardware constraints.

Together, these techniques make transformer training feasible and efficient, even for specialized tasks in proteomics, where both accuracy and scalability are essential.

Having now introduced both RNNs and transformers, we are in a position to compare their strengths and relevance to proteomics. While RNNs provide a natural approach for sequence modeling and have proven useful in tasks involving peptide sequences and mass spectra interpretation, they are limited by their sequential nature and difficulties in learning long-range dependencies. Transformers overcome these limitations through self-attention mechanisms and parallel processing, offering greater scalability and performance, especially on large and complex datasets.

With this foundation in place, we now turn to concrete applications of these architectures in proteomics, highlighting how they are advancing state-of-the-art solutions across a range of biological and analytical problems.

2.4 Applications to proteomics

Deep learning has been rapidly adopted in proteomics, offering significant improvements in tasks that rely on interpreting complex, high-dimensional data such as spectra and protein sequences. Both RNNs and transformers have found diverse applications throughout the proteomics pipeline, from raw data interpretation to protein characterization.

One major application is de novo peptide sequencing, where models attempt to reconstruct peptide sequences directly from tandem mass spectra without relying on sequence databases. Tools such as DeepNovo [38] and PointNovo [59] have demonstrated how deep learning can outperform traditional algorithms by capturing the complex relationship between fragment ions and amino acid sequences.

Retention time prediction has also been improved by deep learning models such as DeepRT [60] and Prosit [37], which accurately predict peptide retention times and fragmentation spectra to support peptide identification in LC-MS/MS workflows. These models help increase identification rates and reduce false positives, especially when combined with spectral libraries.

Deep learning models have also shown great promise in predicting collision CCS values, which provide valuable information on the size and shape of peptide ions in the gas phase. CCS prediction is crucial for IMS-based proteomics, where it adds an additional dimension of separation and enhances peptide identification. Models like DeepCCS [61] and others have demonstrated that neural networks can learn the complex relationships between peptide sequence, charge state, and CCS values with high accuracy. More recent transformer-based models are capable of modeling

CCS in a context-aware fashion, accounting for both linear and nonlinear sequence motifs[62].

Protein property prediction is another growing area. Pretrained language models such as ESM [63] and ProtBERT [64] use transformer architectures trained on millions of protein sequences to learn informative embeddings that can be fine-tuned for downstream tasks like subcellular localization, disorder prediction, and structure modeling.

Generative models based on transformers and variational autoencoders have been applied for peptide design and simulation of synthetic datasets. These models support the generation of novel peptides with desired physicochemical properties or fragmentation profiles, which is particularly valuable in biomarker discovery and targeted proteomics [65][66].

Finally, the integration of multi-omics data using deep learning has opened new directions in systems biology and personalized medicine. Deep neural networks have been proposed for combining proteomics with transcriptomics, genomics, and metabolomics data to construct more comprehensive models of biological systems [67][68].

These applications demonstrate that deep learning is not only enhancing existing workflows in proteomics but also enabling new ones that were previously impractical due to computational or data limitations. The flexibility and scalability of RNNs and especially transformers position them as essential tools in the modern proteomics toolkit.

3
Contents of this thesis

This thesis pursues two main goals: enhancing protein identification through software-level improvements in MS-based proteomics, and explaining a recently observed phenomenon in Ion Mobility experiments. To address these objectives, I developed two independent projects. The first investigates the bimodality observed in a large-scale IMS experiment and introduces a two-valued ion mobility predictor that improves peptide identifications in a DIA setup. The second project presents a re-scoring model designed to improve identification accuracy in DDA data.

The distribution of peptide mobility values exhibits a bimodal pattern—a phenomenon that remains poorly understood within the MS community. In my first project, presented in Chapter 4, I investigate the origin of this bimodal distribution using Molecular Dynamics (MD) simulations. Building on these insights, we develop a machine learning regressor that explicitly models the two distinct mobility populations. This approach enables the generation of more accurate predicted libraries, which in turn enhances peptide identification in DIA experiments conducted on timsTOF instruments.

In DDA experiments, peptide identification is typically performed via sequence database search engines by matching predicted and measured spectra using handcrafted features and a hard-coded scoring function. In my second project, presented in chapter 5, we advance this idea further by removing the intermediate steps of intensity prediction and feature engineering altogether. Instead, we directly model PSMs using an end-to-end deep learning classifier trained to distinguish true from false identifications. This approach leverages modern neural network architectures to learn complex patterns directly from raw input features, enhancing sensitivity and specificity in peptide identification.

Overall, this thesis aims to enhance peptide identification in expression proteomics through improved modeling of PSMs in both DDA and DIA settings. In addition to method development, we leverage physical modeling via molecular dynamics to better understand ion behavior, linking fundamental peptide structure to instrumental observables such as mobility. Each following chapter corresponds to an individual publication. The introduction of each chapter contains information about the publication process and details about my specific contributions to the project. There is also a section called Author contributions on each chapter where the contributions of each team member are described.

Part II

Results and Discussion

4

Bimodal peptide collision cross section distribution reflects two stable conformations in the gas phase

This chapter presents a collaborative study aimed at characterizing protein conformational dynamics through the integration of MD simulations, proteomics data analysis and ML. The project demonstrates how computational techniques can be used synergistically to dissect the structural behavior of proteins and connect physical conformations to observable experimental parameters such as CCS.

The findings from this study are available in a preprint hosted on *bioRxiv* [69], and the manuscript has been submitted to *Nature Communications*.

I contributed primarily to the development and execution of the ML-based analysis pipelines and the proteomics data analysis. My work focused on modeling and interpreting data arising from molecular simulations and ion mobility experiments, with the goal of enhancing the accuracy and interpretability of structural predictions. For a detailed explanation of the contributions of all the authors, see the section *Author contributions* on this chapter.

This study underscores the value of combining ML with physics-based simulations to gain deeper insights into biomolecular structure and function. It exemplifies how methodological innovation—grounded in computational rigor—can yield biologically meaningful interpretations that extend beyond traditional experimental observables. The complete article, as it is online, is shown below.

Bimodal peptide collision cross section distribution reflects two stable conformations in the gas phase

Juan Restrepo[1,#], Daniel Szoelloesi[2,#], Tobias Kiermeyer[1], Christoph Wichmann[1*],
Helmut Grubmüller[2*] and Jürgen Cox[1*]
[1]Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.
[2]Department of Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Science, Am Fassberg 11, 37077 Göttingen, Germany
[#]These authors contributed equally to the publication.
[*]Correspondence: cox@biochem.mpg.de, hgrubmu@gwdg.de,
wichmann@biochem.mpg.de

4.1 Abstract

Recent high throughput applications to shotgun proteomics have shown great benefits of coupling ion mobility spectrometry (IMS) to mass spectrometry. IMS adds a separation dimension by differentiating biomolecules by their size and shape. We (and others) find that the distribution of peptide collision cross section (CCS) is often bimodal, which limits the utility of current machine learning predictions for peptide identification. Molecular dynamics simulations indicate that the peptides in the drift tube can adopt multiple stable conformations and that the two modes correspond to predominantly extended (mostly helical) and more compact (globular and less ordered) conformations. Most peptides have a charge-dependent strong preference for one of the two conformations, while some can adapt both, as evidenced by a simple geometric model of the CCS data. We suggest a novel two-valued CCS predictor allowing for multiple peptide conformations. Its integration into data-independent acquisition proteomics increases identification rates of peptides compared to single-value predictors.

4.2 Introduction

Ion mobility spectrometry[70, 71, 72] (IMS) is a method for separating ionized molecules in the gas phase based on their mobility in a carrier gas. Measured ion mobility values can be converted to rotationally averaged collision cross sections (CCS)[73], which are correlated to the three-dimensional structure of the ionized molecules. Therefore, it can separate molecules by their sizes and conformations in complex samples. Shotgun proteomics[74, 75, 76, 77] involves measuring more than 100,000 different peptides from a single liquid chromatography-mass spectrometry (LC-MS) run[78]. Here, using IMS after liquid chromatography has proven to be beneficial[79, 80, 81, 82] because the separation of co-eluting peptides by their CCS leads to less complex mass spectra and subsequent benefits in their identification[83]. Besides the reduced complexity, the MS features get annotated with their CCS values when LC-MS is coupled to IMS. This additional data dimension can be used to reduce the search space during peptide identification and thereby increase the reliability of identification confidence. However, unlike the sequences, which are known *a priori* and are available in databases for searching, the CCS values of ionized peptides in the gas phase are in general unknown, since the three-dimensional structures of gas-phase peptides are experimentally and computationally hard to obtain, particularly at large scale. Furthermore, it is known that peptides can produce complex ion mobility spectra (mobilograms) with multiple peaks. This can occur due to multiple energetically favorable configurations in the gas phase[84] or other more complex mechanisms[85, 86, 87, 88]. In either case, the mobilogram can serve as a rich source of physical information for identification.

The conformations of ionized peptides in the gas phase have been studied both experimentally and theoretically, the latter mainly by molecular dynamics (MD) simulations. Many studies focused on Alanine-based peptides[89, 90, 91, 92, 93] and have shown that different mobility values represent specific peptide conformations such as alpha helices, hinged helix-coils, globular, and open globular structures. Furthermore, alpha helix stability has been studied in a wide range of experimental conditions[90, 92, 93] as well as its unfolding dynamics[91]. These results not only show that ionized peptides can have various stable conformations, but also prove that MD simulations can offer a toolbox for understanding IM experiments on the atomic level. Nevertheless, the molecular mechanisms leading to the various peptide conformations as well as their interconversion are highly complex and remain only partially understood. Breuker & McLafferty[94] showed that proteins retain water molecules during the electrospray ionization and keep the initial condensed phase structures in the gas phase. However, these water molecules are not detected in a large-scale proteomics experiment. Subsequent studies on the stability of these structures showed that helices tend to unfold into globular structures when the temperature of the ion is increased[85, 95] and also when they pass through a drift tube filled with a buffer gas and an electric field[91].

To develop a physically sound model for CCS values of peptides it is therefore crucial to better understand the underlying physical reasons of the complexity in the IMS mobilograms of peptides and to incorporate these findings into the model. To this end, it is also essential to gain a more fundamental understanding of the molecular scattering processes that give rise to the observed CCS values. Current prediction approaches use either regression on sequence space, as they can be applied to retention time prediction as well[96], or empirically driven models[83, 97, 98] that explicitly parameterize the influence of amino acid positions on size and structure. These approaches usually neglect that a peptide molecule can have more than one CCS value or, more recently[99], allow for several values without representing the actual peptide dynamics and not much impact on identifications. Furthermore, benchmarking predictions on pre-filtered data with single valued CCS, might not be a realistic setting, since the actual information available in the identification process is more complex.

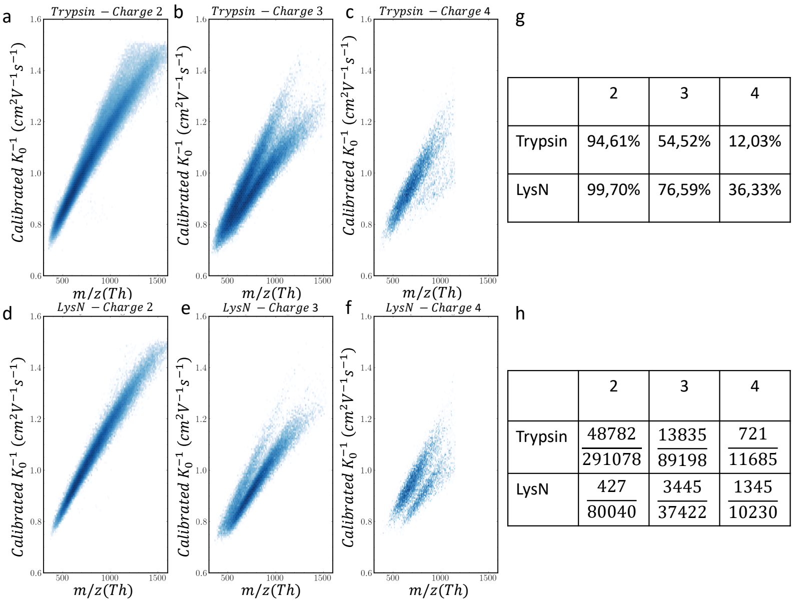
Here we want to reveal the structural origin of the bimodal behavior of ion mobility values observed in peptide populations and find a model for the IMS values of ionized peptides. To this end, we combined analysis of a large-scale LC-IMS-MS/MS shotgun proteomics dataset with molecular dynamics simulations of *in vacuo* folding of representative peptides, as well as of their motion through the diluted gas within the drift tube. Our MD results suggest the two modes correspond to distinct conformations: one globular (compact) and the other helical (extended), with some peptides being able to adopt either state. Our simulations also reveal three different modes of how the peptide interacts with individual gas molecules, which, combined, determine the terminal drift velocity and hence its mobility. To extend these findings beyond computationally expensive MD simulations, we applied further computational and statistical methods. Starting with an idealized scattering model of helices and spheres for qualitative insights, we then developed a geometric model that accurately fits measured CCS values, providing a quantitative description of the bimodal distribution. Lastly, we developed a performance-driven method which focuses on separating the two populations as well as possible to enable the development of a novel, multi-valued CCS prediction method and benchmarked with respect to its impact on the number of identified peptides.

4.3 Results

4.3.1 Bi-modality of peptide collision cross section distribution

We re-analyzed a large LC-IMS-MS/MS shotgun proteomics dataset[100] spanning five species and the three proteases Trypsin, LysC and LysN. Fig. 4.1a-c shows the distribution of peptides in the trypsin-digested dataset in the space of reduced mobility vs. mass-to-charge ratio. Two subpopulations which overlapped with each other are observed with three positive charges over the m/z-1/$K0$ plane as shown in Fig. 4.1b. A fit to the sum of two bivariate normal distributions (Supplementary Fig. 4.1) assigns 55 and 45 percent of data points to the upper and lower cloud, respectively. Separating the data by species (Supplementary Fig. 4.2) shows no significant inter-species differences in the distributions of data points, suggesting that the observed bimodality arises from physio-chemical effects independent of species origin. Examination of the same charge state across proteases in Fig. 4.1 reveals a consistent trend: as the net charge increases, a higher proportion of peptides is observed in the upper population across all enzymes. Fig. 4.1g presents the estimated percentages of peptides in the lower population for all combinations, further corroborating this trend.

Additionally, analysis of the different charge states per protease indicates that, for all charge states, peptides digested with Trypsin and LysC (Supplementary Fig. 4.3cgk) exhibit a stronger preference for the upper population compared to those digested with LysN (Fig. 4.1d-f). By looking at the net charge of the termini (Supplementary Fig. 4.4), it is also clear that peptides digested with LysN and with positively charged amino acids close to the C-terminus tend to be in the upper population while negatively charged ones tend to be in the lower population. A similar behavior but with lower intensity can also be seen for peptides digested with LysC/Trypsin. Peptides can be found in both populations, Fig. 4.1h shows the ratio of such peptides, and even occur in more than two versions, which results in complex 1/K0 spectra (Supplementary Fig. 4.5). Previous studies have examined the link between peptide sequence and the two subpopulations, but, although significant enrichments exist, the effects are small and do not fully explain the bimodality. We therefore have approached the question why some peptides have multiple ion mobilities and how the observed CCS values arise from accumulated individual collisions with gas molecules in the drift tube from first principles using atomistic molecular dynamics (MD) simulations. In particular, to answer the question why some peptides have multiple ion mobilities, we asked (i) can the same peptide adopt multiple conformations in the drift tube environment?, and (ii) does the conformation affect the drift velocity and if so, how? To tackle these problems, we carried out MD simulations with 12 sample peptides selected from the large-scale dataset (Table 4.1).



	2	3	4
Trypsin	94,61%	54,52%	12,03%
LysN	99,70%	76,59%	36,33%

	2	3	4
Trypsin	$\frac{48782}{291078}$	$\frac{13835}{89198}$	$\frac{721}{11685}$
LysN	$\frac{427}{80040}$	$\frac{3445}{37422}$	$\frac{1345}{10230}$

Figure 4.1. Bimodal distribution. a.-f. Distribution of CCS vs, m/z values across proteases Trypsin and LysN and net charges +2, +3 and +4. **g.** Estimated percentages of peptides in the lower population. **h.** Number of peptides identified in both populations for a given charge state and protease.

MD simulations reveal a strong prevalence of globular and helical peptide conformations in vacuum

Conformations *in vacuo* were determined by 'temperature quenching' MD simulations. All simulations were started at a temperature of 600 K, which was subsequently gradually decreased to 305 K over a period of 0.5μs. Each of these simulations was initiated from a fully extended peptide conformation, mimicking high temperature release from the solvent and subsequent conformational quenching in the gas phase, as expected to occur in the experiments. For each of the selected 12 peptides, 1000 such quenching simulations were carried out. The changes in the peptide conformation were characterized through CCS values predicted from simulation frames by the software IMoS, as well as by their helix content.

Figure 4.2a shows two example structures of the P1 peptide, a largely helical one with a relatively large predicted CCS value, and a more globular structure with higher variability, no preferred fold, and with a smaller predicted CCS. Fig. 4.2b shows the evolution of the 1000 simulations for the P1 peptide as the temperature decreases. Initially, high temperature conformations with large CCS dominate, whereas with decreasing temperature two markedly lower CCS populations at 750 and 1000 Å^2 emerge, the distribution of which converges towards the end of the simulations (Fig. 4.2c). Notably, a clear gap between the two main CCS populations is seen at about 850 Å^2, and exchange between these ceases at already 490 K. The two main

name	charge	sequence
P1	+3	DFGYGVEEEEEEAAAGGGVGAGAGGGCGPGGADSS**KPR**
P2	+3	KDLITNIGSGVAAPGAGAPAAAAAPAAAEASE**SK**
P3	+3	PLADHLLAP**TR**
P4	+3	LAGESGSNL**RK**
P5	+4	NQVTCLSVSTDGSVLSG**SHD**ET**VR**
P6	+3	AALEAGAFAVVSTH**WADD**GGAGAVQLADAV**IK**
P7	+4	KALDGQ**NLK**DLLVNFSAAGPAVGAVGAGVGAEGAAEEE**KEEEEAA**
P8	+3	**LLGH**WEEAAH**DLALACK**
P9	+4	**KEIGR**QAALTRNVLEADLILGAIDGISQSVQ**AEA**
P10	+4	**LQENIDSLRSDL**RE**K**
P11.1	+3	**HEQEELHRK**
P11.2	+3	**HEQEELHRK**
P12.1	+3	KDGDY**HVS**ADLTGQANHLAATIGAD**IVKQ**
P12.2	+4	KDGDY**HVS**ADLTGQAN**HLAATIGADIVKQ**

Table 4.1. Table 1. Tested peptide sequences, total charge and charge location. Positively and negatively charged residues are shown as bold blue and red.

conformations shown in Fig. 4.2 a are seen also for other peptides, e.g. P5, as shown by their final CCS distributions (Fig. 4.2d). For comparison, the CCS values measured in the experiment are indicated as red markers. The values predicted from MD structures follow the same trend as the measured CCS, but differ by an offset, in line with the observation by Ewing et al.[101] , according to which the predicted CCS values are about 5% larger than the measured ones.

As can also be seen in Fig. 4.2d, several peptides indeed show multiple CSS values even for the same charge distribution, with P1 as the most incisive example. Of the other peptides, many also show wide CCS distributions, indicating more than one conformation. For P11 and P12, where two different charge distributions were tested (P11.1 vs. P11.2 and P12.1 vs. P12.2), the resulting distributions differ, too, albeit both charge states sample the whole CCS space. This observation suggests that, in addition to the different conformations observed for identical charge states, alternative charge states can add to the multimodality of the observed CCS distibutions.

The example of P1 indicates that α-helices can and do form *in vacuo,* which is plausible due its hydrophobicity, akin to membrane interiors, which also promote the formation of α-helices. For all 12 peptides, Figs. 4.2e and f summarize the abundance of α-helical content along the sequence (Fig. 4.2e) as well as how the predicted CCS relates to the helical content (Fig. 4.2f. Because some of the unrelated peptides show considerable helical content, and nearly all of them some, we conclude that helix formation is a general phenomenon, and may indeed occur for a larger fraction of all measured peptides.

Notably, although the fraction of helical residues generally correlates with the CCS for the larger peptides (Fig. 4.2f), the correlation is much weaker for the smaller peptides. Closer inspection of the respective structures shows that this is because the CCS values of a short helix and a globular fold are rather similar, which also explains why the two populations merge at low peptide mass.

To assess whether or not the observed conformations are kinetically trapped or, rather, in thermal equilibrium, Supplementary Fig. 4.6 shows the potential energy distributions of P1 in its two final conformations during last 40 ns of the quenching simulations at constant temperature 305 K. The distributions are largely overlapping, with an only small difference between the respective average potential energies between the two conformations. This indicates that the two conformations are indeed close to thermal equilibrium and, therefore, their final distributions are expected to be rather insensitive to the chosen cooling rates, the precise value of which in the experiments is unknown.

MD simulations of peptides within the drift tube environment show that conformational differences suffice for drift velocity bimodality

Are the structural differences predicted by the MD quenching simulations large enough to explain the measured bimodal ion mobility distributions? To answer this question, we performed fully atomistic MD simulations of P1, for which the globular and helical conformations are well separated, within the drift tube environment including the electric field that accelerates the peptide (see methods) as well as the opposing air resistance due to collisions with the gas molecules. Here we did not want to resort to established structure-based CCS estimates, because it is unknown (a) to what extent the peptide conformation is changed due to the 'bombardment' by the gas molecules within the drift tube and (b) what the nature of the collisions is that ultimately determine the effective CCS. Our new type of MD simulations served to also address these questions.

From the quenching simulations seven different globular and seven helical conformations were selected; each of these was placed within 10 different boxes of 10^6 nm^3 each, filled with 2.7 mbar air (51 N_2 and 13 O_2 molecules) with different random gas positions and velocities, resulting in a total of 140 simulations. In each of the simulations, the peptides started at rest, and – similar to the experiment, – an electric field of 20 V/cm was applied to accelerate the peptide while the center of mass of the gas molecules was kept stationary. Due to the electric field, the peptide gradually accelerated while colliding with the gas molecules. In order to maintain the gas temperature but do not perturb peptide velocities, only the gas was coupled to a heat bath (see Online methods). Visual inspection of the simulations with high temporal resolution (1 frame/ps) revealed three main collision types sketched in Fig. 4.3a and visualized in the Supplementary movies 1-4(see publication online). The first type, the expected one, is mostly elastic, with a very short interaction time; second, and unexpectedly, we observed adsorption with subsequent reemission, where the gas molecule spent an extended time span (up to 3.5 ns) on the surface of the peptide; third, swing-by events, during which, in contrast to a collision, the protein and gas interact attractively through Lennard-Jones forces and thus also change velocities. This third type of collision was also unexpected due to the weakness of the Lennard-Jones interactions.

A more quantitative analysis of the over 1000 collision events observed in our simulations is shown in Fig. 4.3b. Here, each dot represents one of the collision events, separated according to the nearest approach ('scattering parameter'), duration of the event, and velocity change (color). The three types of collisions form clusters, which are, however, not clearly separated from each other, and rather blend into each

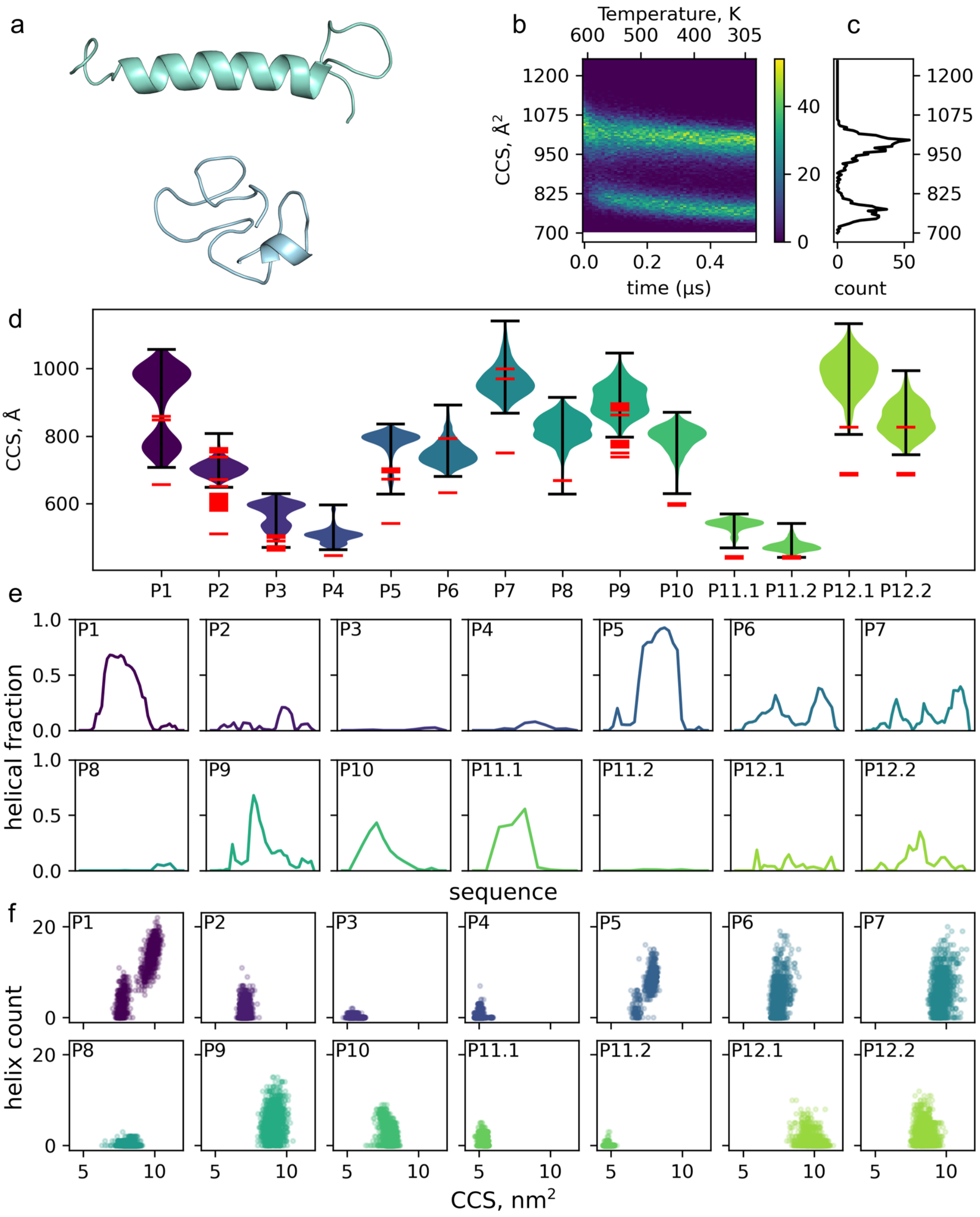


Figure 4.2. Peptide conformation in vacuum. a. Example conformations of the two major folds. **b.** Evolution of the CCS over the quenching simulation and **c.** CCS distribution of the final conformations. **d.** Violin plot of CCS for all simulated peptides. Peptide P11 and P12 was tested with two different charge distributions indicated by the label. Red line segments indicate available measurement data (peak positions). **e.** The fraction of helical residue along the tested sequence compared to the 1000 quenching simulations using the final conformation. **f.** Count of helical residues versus predicted CCS using the final conformation of the quenching simulations.

other, displaying a gradual range of these properties. Overall, the elastic collisions are short-lived, as are the swing-by events, but differ by both much shorter nearest approach and larger change of velocity (mainly direction). In contrast, due to the weaker intermolecular interaction, the velocity change of the swing-by events is smaller and scatters over a broader range. The adsorption/re-emission events are characterized by the closest approach, naturally, and by a broad spectrum of residence times ranging from 2 up to several 100 nanoseconds. Fig. 4.3c shows the total numbers of observed collisions for the globular and helical conformers; in line with its larger estimated CCS, significantly more collisions are seen for the helical conformations. An analysis of the root mean squared deviation (RMSD) of the simulated structures during and after the collisions (0.07±0.03 nm versus 0.12±0.04 nm, mean and standard deviation for the globular and helical conformations, respectively) shows that the impact of the gas molecules leaves the peptide structures largely unaffected and in particular does not trigger conformational transitions between helical and globular structures.

Next, we quantified the acceleration of the P1 peptide during the simulation by the applied (static) electric field against the increasing air resistance for the two conformations (Fig. 4.3d,e). Due to the relatively few collisions experienced within the highly diluted gas during each simulation, the individual traces (transparent lines) show considerable 'Brownian motion' scatter. Yet, the average velocity over 70 trajectories each (red solid lines) is well converged and follows the analytical solution of the Newtonian equation of motion for an accelerated object with air resistance proportional to its squared velocity (black lines, see Supplement and methods); this analytical solution was fitted to the average velocity with the ratio between electric field strength and air resistance (γ) as the only fit parameter. A rapid initial velocity increase is seen, with a rate determined only by the peptide mass, charge and electric field strength (red dashed lines), and subsequently with decreasing acceleration towards a terminal drift velocity measured in the experiment. In line with the smaller number of collisions seen for the globular conformation (Fig. 4.3c), its terminal velocity (62.12±2.74 m/s), determined from the analytical fit at infinite time, is markedly larger than that of the the helical conformation (54.21±2.09 m/s).

For comparison, we estimated the drift velocities in the experiment from $K0$ values given the conditions used in the simulations (pressure: 2.7 mbar, temperature: 305 K and electric field: 20 V/cm) and obtained 78.31±1.75 m/s and 59.79±0.65 m/s for the two measured ion mobility peaks (peak value ± half width). Because particularly the pressure and density within the drift tube cannot be measured very accurately, this deviation is not unexpected, and one would therefore assume that the calculated drift velocity differs from the one estimated from the experiment by a common factor. Indeed correcting, e.g., the pressure to 3.2 mbar yields an estimate of 65.4 and 50.0 m/s, respectively, which lies very close to the values from MD. The main result here is that the extended conformation indeed shows a significantly slower drift velocity (and a correspondingly larger CCS) than the more compact conformation. In particular, the difference is large enough to explain the bimodal drift velocity distribution.

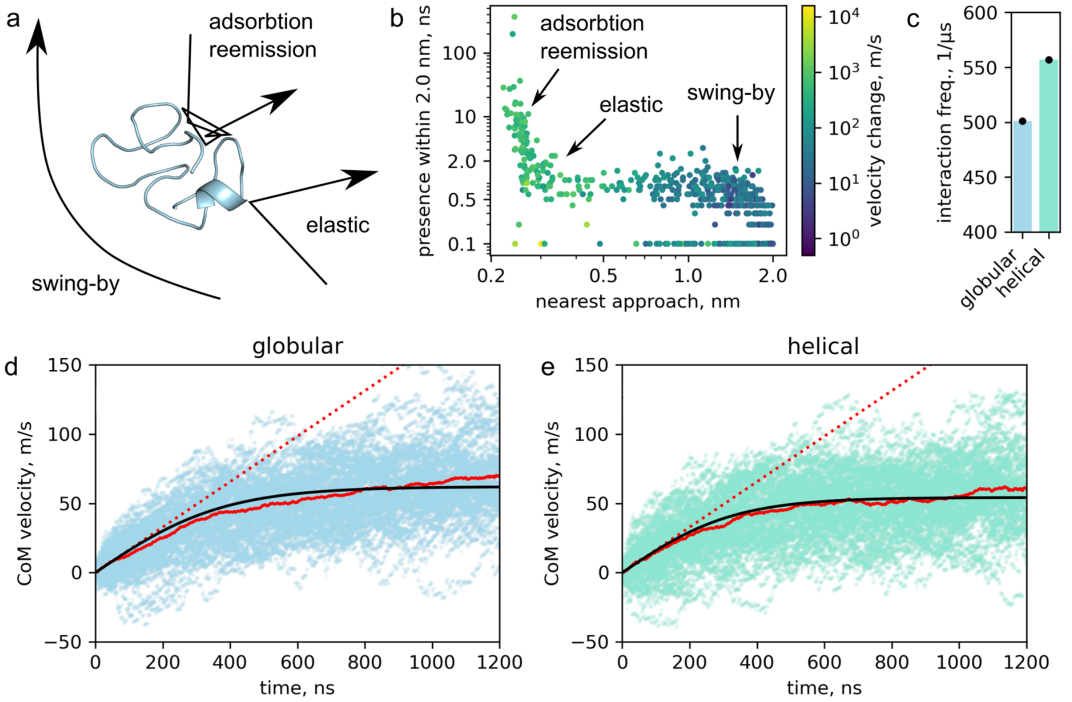


Figure 4.3. Collisions and simulated drift velocity. a. Scheme of the three main collision types. **b.** Collision events are tested for nearest proximity, time spent within 2 nm and the change of the velocity before and after the event. The location of the tree collision types are highlighted **c.** The number of interactions/collisions averaged for the globular and helical folds in the drift tube simulations. Error bars are highlighted as black bars, within the symbol size **d**. and **e.** Drift velocity of the individual simulations trajectories for globular (light blue) and helical (light green) conformations. Average drift velocity is indicated by a red line along with a fitted curve (black line).

4.3.2 Geometry-inspired approximation explains the bimodality at large scale

Our MD simulations suggest that globular and helical structures are stable in the drift tube environment, consistent with experimental results for poly-alanine peptides[84]. To further address the question if these two types of conformations are sufficient to explain the bimodal large-scale proteomics data, and since MD simulations for all peptides measured in the proteomics data are computationally quite demanding, we developed approximate treatments.

The, so called, geometric fit achieves a quantitative description of the large-scale proteomics data by statistically describing the dataset as a combination of two overlapping populations. The fit uses a joint model over all charge states where the CCS-mass distribution for each charge is parameterized as the sum of two independent densities, one for the globular and one for the helical population (see Online Methods and Fig. 4.4b). The helical population follows a linear CCS-mass relationship, as peptides grow primarily through helix extension. The globular population exhibits a power law dependence, CCS ~ mass$^{2/3}$, reflecting uniform volume growth with added mass. Fig. 4.5b presents the geometric fit for charge state +3, while Supplementary Fig. 4.7 provides the fit results across all charge states, along with the associated error. This simple yet effective model closely aligns with the data, supporting the hypothesis that these two conformations indeed underlie and explain the observed bimodal distribution.

The intercept with the CCS axis (0.59 nm^2) represents the charge-dependent contribution to the CCS corresponding to the value a massless unit charged peptide would exhibit. The helical slope (1.34 nm^{2}kg^{-1}mol^{-1}) reflects the average CCS growth rate for a helical peptide as more amino acids are added. This procedure also provides the probability of a particular peptide belonging to either population. After calculating these probabilities, we found that for charge states +2, +3, and +4, the percentage of peptides in the upper (helical) population is approximately 4%, 20%, and 81%, respectively. Despite the good agreement, systematic errors remain as indicated by the non-random distribution of residuals (see Supplementary Fig4.7). Potential sources include deviations from sum of Gaussians of the data points and, particularly at lower masses, deviations of peptide conformations from the assumed globular or helical shapes.

Next, we determined whether Monte Carlo-based estimation of CCS values of ideal spheres and helices using IMoS[102] can qualitatively describe the dataset (see Fig. 4.4a). Helical structures were obtained from AlphaFold2 predictions of experimentally measured peptides[103] and compared to idealized polyalanine, polyleucine, and poly-tryptophan helices of varying lengths generated using PyMOL. Lacking well-defined globular candidates, we generated them via (i) transforming helices using the Fibonacci sphere method and (ii) distributing atoms on a spherical shell at van der Waals separations (see Online Methods). These structures were then evaluated in IMoS under frozen geometries, excluding partial charges. While this approach does not aim to be fully accurate, it successfully distinguishes the characteristic CCS trends of helical and globular structures. As shown in Fig. 4.5a, the simulated CCS values follow distinct scaling laws, with globular structures conforming to a power-law and helical structures with a linear trend. This clear separation supports our hypothesis and confirms that this simple geometric model qualitatively describes the observations. Quantitative deviations are seen for the exponent (2/3 from geometric considerations vs. 0.41 determined from data) as well as a systematic positive vertical offset, which both are likely due to the neglect of intermolecular interactions such as Coulomb and van der Waals forces; as seen in the atomistic drift tube simulations, these interactions affect the scattering processes and thus also the CCS.

Finally, we performed a solely data-driven fit, here referred to as empirical fit, designed to optimally separate the two populations for use in CCS prediction. In this approach, we divided the (CCS, Mass) space into bins, smoothed the data using a Gaussian kernel, and assumed that each slice at a constant CCS could be modeled as a mixture of two Gaussians (see Online Methods and Fig. 4.4c). Selected slices for charge state +4 are shown in Supplementary Fig. 4.8. Fig. 4.5c shows the mean of both Gaussians for all the transversal slices of the charge state +3 dataset (yellow dots for the left Gaussian, red dots for the right Gaussian), overlaid on the measured data. The means accurately trace the regions of highest density for each population and closely follow linear and power-law trends, consistent with the expected structural scaling behaviors. This purely data-driven method yields probability distributions per population and per slice without relying on prior assumptions about peptide geometry. Importantly, it allows us to assign labels to peptides based on the inferred probability distributions using their CCS, charge and mass across the whole dataset. These labels allow us to train per-population regressors for CCS prediction.

4.3.3 Two-valued machine learning prediction improves identification of peptides in proteomics

Existing mobility predictors are typically trained to predict the mobility of a peptide's most intense feature, overlooking the fact that many peptides exhibit two distinct, well-defined mobility values due to the existence of two stable conformations. This oversight introduces stochasticity and may reduce peptide identification rates by predicting the 'wrong' mode for a peptide in a given dataset. To address this issue, we divided the training set into two clusters by assigning each peptide a probability of belonging to either cluster using the empirical and geometric fit described in the previous section and selecting the most likely one (see Methods). We then trained a bidirectional recurrent neural network (RNN) on the encoded sequence and charge for each cluster across all charge states. Similarly, we derived sequence-based features, combined them with metadata, and trained a XGBoost regressor for each stable conformation. As a baseline, we trained both models on the unseparated dataset, representing the conventional approach used in prior studies. To benchmark the performance of these models, we employed two approaches: evaluating prediction error on an independent test set and assessing peptide identification rates in a data-independent acquisition (DIA) experiment using predicted libraries.

We tested the models on an independent dataset from the ProteomeTools project and measured the relative error with respect to the most intense value within each cluster. The results for charge 3, shown in Fig. 4.6a, reveal that the RNN with empirical labeling outperforms other models in the most bi-modal case, as expected from its optimal separation of peptide populations. The RNN with geometrical-fitting labeling ranked a close second, showing that it also successfully learned the bimodality of the data. The baseline RNN trained on the unseparatrd dataset performs similarly to

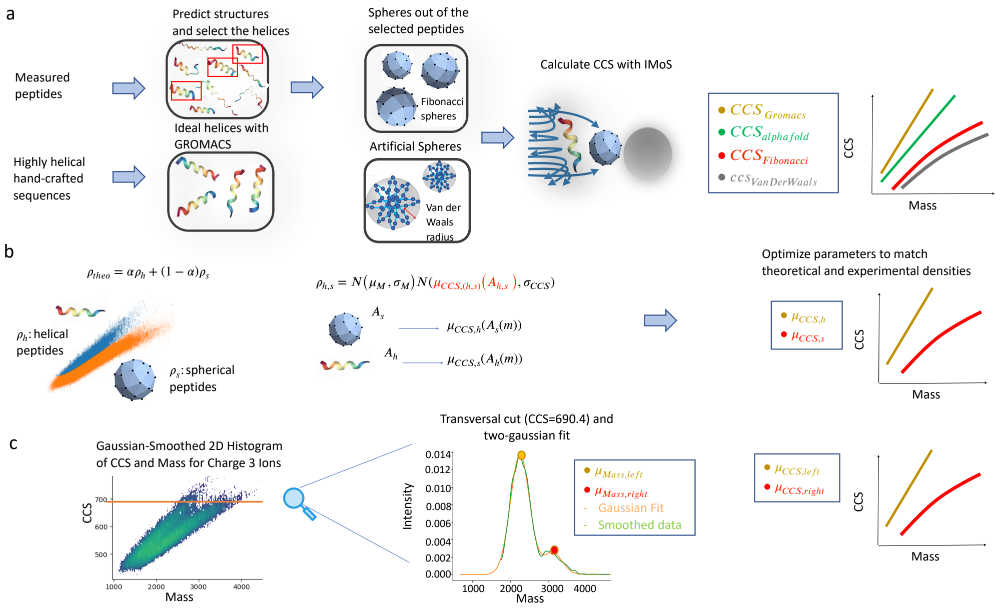

a
Predict structures and select the helices
Spheres out of the selected peptides
Measured peptides
Fibonacci spheres
Calculate CCS with IMoS
Ideal helices with GROMACS
Artificial Spheres
Highly helical hand-crafted sequences
Van der Waals radius
$CCS_{Gromacs}$
$CCS_{alphafold}$
$CCS_{Fibonacci}$
$CCS_{VanDerWaals}$
CCS
Mass
b
$\rho_{theo} = \alpha\rho_h + (1-\alpha)\rho_s$
$\rho_{h,s} = N(\mu_M, \sigma_M)N(\mu_{CCS,(h,s)}(A_{h,s}), \sigma_{CCS})$
ρ_h: helical peptides
ρ_s: spherical peptides
A_s
$\mu_{CCS,h}(A_s(m))$
A_h
$\mu_{CCS,s}(A_h(m))$
Optimize parameters to match theoretical and experimental densities
$\mu_{CCS,h}$
$\mu_{CCS,s}$
CCS
Mass
c
Gaussian-Smoothed 2D Histogram of CCS and Mass for Charge 3 Ions
CCS
700
600
500
1000
2000
3000
4000
Mass
Transversal cut (CCS=690.4) and two-gaussian fit
Intensity
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1000
2000
3000
4000
Mass
$\mu_{Mass,left}$
$\mu_{Mass,right}$
Gaussian Fit
Smoothed data
$\mu_{CCS,left}$
$\mu_{CCS,right}$
CCS
Mass

Figure 4.4. Fitting Models. a. Geometric scattering: Measured peptides undergo structure prediction with AlphaFold2, from which helical structures are selected. Spherical structures are generated from these helices using the Fibonacci sphere procedure. Hand-crafted sequences are converted into ideal helices with PyMOL and into spheres based on Van der Waals radii. CCS is then computed using IMoS for different structural models. **b. Geometric fit:** The large CCS vs mass dataset is modeled as a weighted sum of two Gaussian-distributed conformational states: helical peptides (ρ_h) and spherical peptides (ρ_s). The center of each Gaussian depends on the projected area of the peptide structure, which scales with mass. Parameters are optimized to best match theoretical and experimental densities. **c. Empirical fit:** A Gaussian-smoothed 2D histogram of experimental CCS vs. mass values is analyzed. A transversal cut at a fixed CCS (690.4 Å^2) is shown, where a two-Gaussian fit distinguishes different conformational states. By repeating this procedure for all the CCS values while saving the fitted means per-population trend lines are generated.

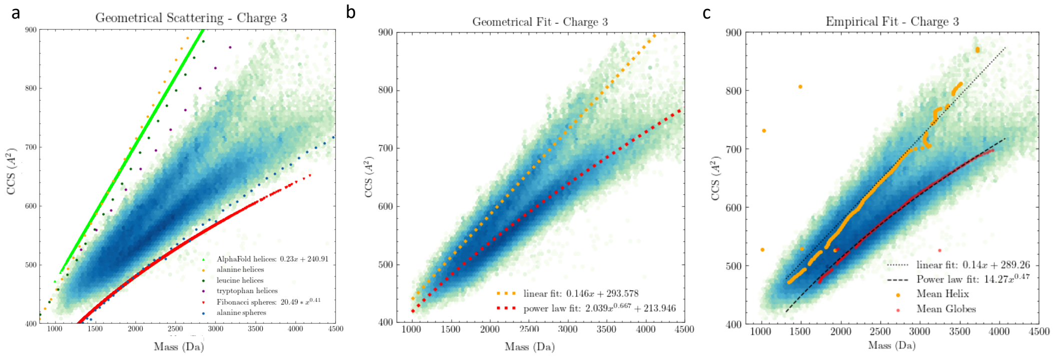


Figure 4.5. Comparison of CCS fitting approaches for charge state 3. a. Geometric Fit: A direct mathematical fit is applied to the experimental CCS vs. mass distribution. Both a linear fit (dotted red line) and a power-law fit (dashed orange line) are used to describe the trend in the data. **b. Geometric Scattering:** CCS predictions are derived from structural models, including AlphaFold2 helices, ideal helices, and spherical models (Fibonacci and Van der Waals spheres). These models define theoretical upper and lower CCS bounds (solid green and red lines), providing a structural basis for understanding the CCS-mass relationship. **c. Empiric Fit:** Experimentally derived trend lines are extracted by identifying the mean CCS values of helical peptides (orange) and globular peptides (red). A linear and a power-law fit are applied to characterize the observed experimental trends.

the XGBoost models trained with labeled data in a basic feature space derived from amino acid counts, clearly exhibiting the importance of the proper labeling. Notably, the poor performance of the XGBoost model for charge state +3 in the unlabeled case highlights the limited flexibility of the feature space, as this charge state is particularly challenging to predict.

Lastly, we used spectral libraries predicted with DeepMass[104], predicted the reduced mobility of the existing sequences and tested them in a data-independent acquisition (DIA) experiment. Fig. 4.6b shows the number of identified peptides using libraries generated by different models. The RNN with data separated by the empirical fit clearly outperformed all other approaches, followed by XGBoost with empirical fit labeling and XGBoost with geometric fit labeling. This result underscores that for mobility prediction, the separation of peptide populations is more critical than the choice of model itself. Interestingly, omitting mobility values ranks fourth, potentially because MaxQuant bypasses mobility filtering in such cases. However, this result also demonstrates that poor mobility predictors can severely impact performance. The RNN with geometrical fit labeling and the baseline RNN without labeling showed similar performance, followed by the baseline XGBoost model as the least effective.

4.4 Discussion

Our combined approach involving ion mobility spectrometry measurements, atomistic molecular dynamics simulations, and geometric modelling, revealed that the bimodality in peptide CCS data as it is produced in proteomics data arises mainly from conformational heterogeneity of a larger peptide population, and specifically from the presence of globular and helical conformations. Our molecular dynamics

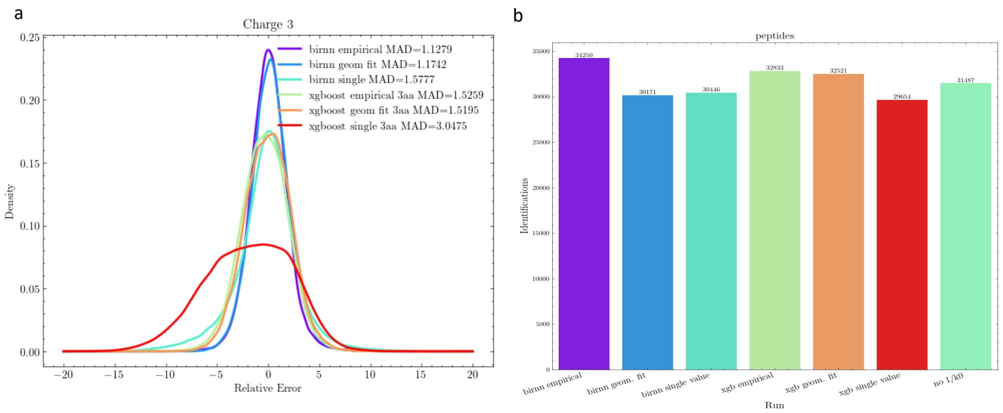


Figure 4.6. Benchmark of mobility predictors. a. Relative Error in Test Set: Comparison of two machine learning models (Bi-RNN and XGBoost) using three different peptide-labeling methods (empirical fitting, geometrical fitting, and baseline with no labeling). The relative error of their predictions is evaluated on the ProteomeTools dataset. **b.** Peptide identifications: The same models and labeling methods are assessed based on the number of peptide identifications in a spectral library used for a DIA run with MaxDIA.

simulations of selected peptides in gas phase show a strong prevalence for these two configuration types for peptides. Full simulations of peptides mimicking the drift tube experiments show that the differences in drift velocities between compact and extended conformations suffice to explain the observed bimodal drift velocity distribution. We found not only elastic collision events, as one might have expected, but also adsorption/re-emission and swing-by events, which, combined, contribute to the observed CSS. Accordingly, our simulations can serve to refine CSS estimates. Based on our simulations, we have successfully derived more approximate yet efficient models which can be successfully applied to the large-scale data. Explicit inclusion of the bi-modality in machine learning-based CCS prediction turned out to markedly improve prediction accuracy.

4.5 Methods

4.5.1 Data download and processing

We downloaded and reprocessed the dataset described in Meier et al.[100] using MaxQuant[105] v2.6.6.0. The search engine Andromeda[106] was used for peptide identification by matching the measured spectra to theoretical spectra generated via in-silico digestion of reference proteomes with specific enzymes (trypsin, LysC, or LysN). Cysteine carbamidomethylation was set as a fixed modification, while oxidation of methionine and protein N-terminal acetylation were set as variable modifications. Additionally, a list of 245 potential contaminants was included in the search. The FASTA files of the reference proteomes, including isoforms, were downloaded from UniProt (release 09/2023) and contained the following: Homo sapiens (103,830 proteins), Saccharomyces cerevisiae (6,091 proteins), Drosophila melanogaster (23,543 proteins), Escherichia coli (4,415 proteins), and Caenorhabditis elegans (28,540 proteins). Following the original publication, for the five-species dataset, the maximum mass tolerances were set to 20 ppm for precursors and 40 ppm for fragment ions. Each set of synthetic peptides was analyzed in independent MaxQuant runs, with libraries generated in silico by tryptic digestion of the human proteome. The additional HeLa dataset was processed as outlined in Meier et al. [107]. The 'TIMS half width' was set to 4, and the 'TIMS mass resolution' to 32,000. The maximum mass tolerance for precursors was set to 70 ppm, for fragments to 35 ppm, and for precursors after recalibration to 20 ppm. The diaPASEF dataset was analyzed, as well, with MaxQuant v.2.6.6.0 using spectral libraries predicted by DeepMass[104]. The mobility values were predicted with the different regression models and added to the libraries manually.

4.5.2 Analysis of the MaxQuant output

The MaxQuant output was analyzed using Python 3.7.11 with the NumPy, Pandas, SciPy, and Matplotlib libraries. We filtered out decoy peptides, potential contaminants, features with null intensity, and peptides identified with only one positive charge. To integrate all the different MaxQuant runs into a single dataset, systematic offsets were corrected through a machine-learning-based approach. Specifically, we designated the larger HeLa dataset as the master run, trained a recurrent neural network on its most intense feature per precursor, and predicted the mobility values for the most intense features of precursors in the other experiments. The difference between the predicted and measured values was then calculated, grouped by raw file, and summarized as a median correction factor for each file. This approach remained robust even when overlap with the master run was limited or nonexistent, such as when LysN was used as the protease.

4.5.3 Geometric fit

For the geometrical fit, we constructed 2D histograms in the (CCS, Mass) space for each charge state, ρ_m. We modeled the total density as the sum of two bi-dimensional Gaussian distributions with relative abundances, $\rho_{th} = \alpha\rho_h + (1-\alpha)\rho_s$, where α is the abundance of the helical population and ρ_h, ρ_c are gaussian distributions representing the helical and spherical populations, $\rho_{h,s} = \frac{1}{2\pi\sigma_m\sigma_{CCS}} e^{-\frac{(m-m_0)^2}{2\sigma_m^2}} e^{-\frac{\left(CCS-CCS_{0,(h,s)}\right)^2}{2\sigma_{CCS}^2}}$. The center of each Gaussian in the CCS dimension was assumed to follow a linear relationship with the projected area, which itself is determined by volume and, consequently, mass, where $CCS_{0,(h,s)} = \lambda A_{proj,(h,s)} + b$. For the lower population, we assumed a spherical geometry for the projected area, $A_{proj,s} = 4\pi r^2$, $r = \left(\frac{3}{4}\pi V\right)^{\frac{1}{3}}$, $V = \frac{m}{density}$, while for the upper population, a cylindrical shape was used, $A_{proj,c} = \frac{(\pi r_c^2 + \pi r_c L)}{2}$, $L = \frac{V}{\pi r_c^2}$, $V = \frac{m}{density}$. The density parameter, initialized to 1000 kg/(mol nm^3), was optimized during the fitting process through a scaling factor. Gaussian widths were optimized but shared across conformations, as were the parameters for the linear model of the center in the CCS dimension. The relative abundance of each Gaussian was modeled as a linear function, with parameters optimized across populations, $\alpha = \gamma m + \beta$. Additionally, the cylinder radius r_c was treated as a free parameter. Once fitting was complete, we computed the probability of each point belonging to the spherical population and labeled it as spherical when this probability was ≥ 0.5.

4.5.4 Empirical fit

For the empirical fitting, we calculated the conditional probability density P(CCS | Mass) for each charge state and performed transversal slicing at constant CCS values. Each slice was smoothed using a Gaussian kernel with sigma=1 Da for net charge two and sigma=4 Da for net charges three and four. Then we fitted a model consisting of the sum of two Gaussians to the resulting distribution. For charge two, Gaussian heights were constrained to (0, 1.0), means to (900, 3200), widths to values greater than 50. For charge +3, the widths were allowed to differ but retained the same minimum threshold, the means were constrained to (1000, 4000) and only slices with CCS<700 were considered as the low point density above this level made the fitting unstable. For charge +4, the mean range was restricted to (1500, 4500), the maximum CCS for the left population was 900 and for the right one 800. Initial parameters for the fitting process were determined using a peak-finding algorithm with thresholds set to a minimum height of 0.0005 and minimum prominence of 0.0001. The two most prominent peaks were used as starting points; if only one peak was detected, it was duplicated, and slices without peaks were skipped. To reduce the noise, an ad-hoc condition was applied, that the fitted means had to differ at least by 200 units. The fitted parameters were sorted based on their means to construct probability density functions for each population within each CCS bin. The left and right Gaussian means were extracted per slice and fitted using a linear model for the left population and a power-law model for the right. Based on these probability density functions; points were assigned to the population with the highest probability.

4.5.5 Geometric scattering

For the geometric scattering we generated ideal spherical and helical peptides. To generate idealized spherical peptide models we used two methods, the Fibonacci sphere method and our own algorithm based on placing atoms on a spherical shell separated by the Van der Waals radius. First we utilized Fibonacci spheres, a mathematical approach for evenly distributing N points on the surface of a sphere. This method allowed for the sequential placement of protein atoms on the spherical surface, ensuring uniform distribution. To construct non-hollow spherical models, we generated 4 concentric Fibonacci spheres with progressively smaller radii. The radii were defined based on the framework provided by[108], which describes the minimal radius of a spherical protein that contains a given mass. This approach enabled the systematic construction of densely packed ideal spherical protein structures. We calculated the spherical conformations for a total of 60496 peptides, comprising 44093 peptides with a charge of +2, 14665 peptides with a charge of +3, and 1738 peptides with a charge of +4. We adapted the radii to smaller values as has been used in the referenceby[108], 80% of it for charge +3, +4 and 70% for charge +2 which can be attributed to the distinct properties of the gaseous environment, where the absence of solvent effects and amplified electrostatic repulsion due to the lack of dielectric screening influence the conformation. In the second method, we considered spheres with radii ranging from 0.5 to 1.1 nm and assumed a fixed density of 1000 kg/m^3. We then determined a dense spherical grid with equal angular steps where the largest separation (at the equatorial) is determined by half of the hydrogen van der Waals radius. Finally, we added atoms to the grid points from a list of atom names representing the composition of alanine if it is not overlapping with already placed atoms. The resulting atom names and coordinates were used to construct a PDB file. To generate the ideal helical peptides, we employed AlphaFold2[103]and PyMOL (Version 3.0, Schrödinger, LLC). When using AlphaFold2 we predicted the solution structure of the measured peptides. Subsequently, we analyzed the secondary structures of these peptides and selected those that exhibited a consistent alpha-helical structure along the peptide backbone. We detected helical conformations for a total of 60496 peptides, comprising 44093 peptides with a charge of +2, 14665 peptides with a charge of +3, and 1738 peptides with a charge of +4. For comparison, we also included helical structures with ideal dihedral angles generated using PyMOL for sequences ranging from 7 to 40 amino acids with high helical propensity, such as alanine-, leucine-, and tryptophan-based peptides. The CCS values of the peptides were calculated using the Ion Mobility Spectrometry Suite (IMoS) software, version 1.10. Default parameters were employed, with the exception of the pressure, which was set to 270 mbar, and the temperature, which was adjusted to 305 K, and the respective charges to replicate the experimental conditions. The CCS values were determined using the Trajectory Method Lennard-Jones (TMLJ) approach, which calculates the momentum exchange between the buffer gas and the peptide by simulating individual trajectories of gas molecules. This method also accounts for interaction potentials between the buffer gas and the molecule. The TMLJ method employs a 4-6-12 potential with optimized Lennard-Jones parameters, making it the gold standard for accuracy in CCS calculations (source IMoS).

MD quenching simulations

We performed fully atomistic molecular dynamics temperature quenching simulations *in vacuo* to sample and predict the conformation of the measured peptides. As unbiased starting structures, fully extended conformations were generated using Pymol (Version 3.0 Schrödinger, LLC.). Although the total charges of the peptides are known, in many cases their distributions are ambiguous. To avoid this uncertainty, sequences were selected where the number of arginine, histidine, and lysine residues together with the charged N-terminal sums up to the expected total charge, such that this charge distribution is uniquely specified. In two other cases, the two most plausible different charge states (among the many other combinatorial possibilities) were used as shown in Table 4.1. Aspartate and glutamate residues we usually protonated and therefore, neutral. The extended peptides were placed in a cubic box with a side length of 100 nm. Energy minimization (steep integrator, 3000 steps) and a four step equilibration was performed with increasing time steps: 0.1 fs, 0.5 fs and 1 fs in NVT and lastly in an NPT ensemble. The resulting structures were used for the quenching simulations where a simulated annealing temperature quench was applied to allow the peptides finding their equilibrium conformation. Specifically, the initial temperature was 600 K and maintained for 10 ns, then linearly and slowly decreased to 305 K over a time period of 500 ns and subsequently kept at 305 K for another 40 ns to facilitate thermal equilibration. Due to the absence of solvent and the presence of a net charge, we changed the simulation parameters relative to those typically used for conventional simulations as follows. i) Double precision compilation of GROMACS (version 2023.4 10.1016/j.softx.2015.06.001) was necessary and a 1 fs integration time step; ii) cutoff distances were set to 30 nm, such that all peptide atoms interacted with all other atoms explicitly via Coulomb and Lenard-Jones forces; accordingly, a long neighbor search interval (1 ns) was used, and (iii) no Particle Mesh Ewald (PME) method was required, which would fail due to the total net charge of the system; iv) no pressure coupling was applied. The system was periodic, but the center of mass was kept at the center of the simulation box. The temperature was controlled by the V-rescale algorithm[109]. All simulations were performed with the CHARMM36m force field[110] with added oxygen and nitrogen molecule parameters adapted from Wang et al.[111] for the later drift tube simulations. The quenching procedure was repeated 1000 times for every peptide with starting velocities chosen randomly and different in each case from a Boltzmann distribution. All atomic positions were saved every 10 ns.

MD drift tube simulations

From the quenching simulations of peptide P1, 7 helical and 7 globular conformations were selected as starting structures to simulate the full drift tube environment. As in the experiment, an air mixture was used as the inert gas at 2.7 mbar (the approximate pressure estimated in the experiments) and 305 K, which translated into 51 N_2 and 13 O_2 molecules in the cubic simulation box of 100 nm side length. This 'drift tube' simulations box was initially created and simulated without the peptide for 100 ns at 305 K to obtain an equilibrated system, to which subsequently the peptide structures were inserted. Starting velocities were taken from the appropriate previous simulations. In each drift tube simulation, an electric field of 20 V/cm parallel to the x-axis was applied.

The aim of this simulation was to test the effect of conformation on the velocity increase and the terminal velocity of the peptide that results from the balance between the force exerted by the electric field and the collisions with the gas molecules. Therefore, center of mass motion removal was applied only to the air molecules in the box (once every ns), and not to the moving peptide. To account for the need to maintain the temperature of the air while not interfering with the velocities of the peptide atoms, we used V-rescale temperature coupling separately for the gas (with a coupling constant $\tau = 5$ ps) and for the peptide ($\tau = 1$ ps). The extremely long coupling time for the peptide ensured a neglible effect of the heat bath. Temperature coupling of only a subset of the simulated atoms is currently not possible with GROMACS. As for the quenching simulations, large cutoffs were used without PME, but due to the fast-moving gas molecules the neighbor search was performed for each integration step (1 fs). Atomic coordinates and velocities were recorded every nanosecond. Each of the 14 peptide conformations was simulated 10 times independently, each starting with a different set of random positions and velocities of the air molecules. In order to obtain the equilibrium drift velocity (v_{d}) from these 10 MD simulations for each peptide, the analytical solution of the Newtonian equation of motion for an accelerated object with air resistance proportional to its squared velocity

$v_d(t) = \sqrt{\frac{qE}{\gamma}} \tanh\left(\frac{t\sqrt{qE\gamma}}{m}\right)$,

was fitted to the time-dependent average velocity obtained from the simulations, with the ratio between electric field strength and air resistance (γ) as the only fit parameter. Here, τ is time, q is the peptide charge (+3), E is the electric field strength (Vm), m is the mass of the peptide (kg/mol), and γ is defined as

$$\gamma = \frac{1}{2}\rho_{Air} A C_d$$

,

with gas density ρ_{Air}, surface area A of the peptide, and shape parameter C_{d}. The air resistance γ includes both shape and surface area differences between peptide conformations, lumped up into one single fit parameter.

4.5.6 Electronic structure analysis with Gaussian

We analyzed the energy of a peptide for multiple runs obtained from molecular simulations, which display both globular and helical conformations, to determine whether there is an energetic disparity between these secondary structures. Using Gaussian software, we performed geometry optimizations by applying density functional theory (DFT) with a ωB97X-D functional and the 6-31G(d) basis set. To reduce computational costs, we used the NoSymm keyword to prevent molecular reorientation and geom=connectivity to explicitly define the molecular connectivity, a commonly used approach for such optimizations. Additionally, we applied the SCF=NoVarAcc option to accelerate the convergence behavior of the self-consistent field (SCF) calculations. The optimizations did not converge to the default convergence criteria of Delta E = 10^{-6} atomic units but sufficient to enable comparison between the two conformations. During the optimization process, the secondary structure for the different runs was maintained.

4.5.7 Machine learning predictions

For each considered separation, we trained a bi-directional recurrent neural network (RNN) and an XGBoost[112] regression model. As a baseline, we also trained both models on the most intense feature per precursor without applying separation. For the deep learning model, peptide sequences were encoded to include modifications, resulting in 26 unique classes. The encoded sequences were padded to form matrices with 66 columns, and the net charge was appended as an additional column. The dataset was split into training (90%) and validation (10%) subsets, with reproducibility ensured by setting a fixed random seed (42). The encoded training sequences were passed through an embedding layer connected to two bi-directional LSTM layers, each layer with 128 units and a dropout probability of 0.5. Global pooling was applied along the sequence dimension after the final LSTM layer to ensure consistent shape across instances. The charge value was concatenated to the hidden state and fed into a fully connected layer with 128 neurons and a dropout probability of 0.4. ReLU activation was applied before the final output. The model was trained for 200 epochs with a batch size of 64 using an inverse square root learning rate scheduler (normalization factor = 1056) with a warmup phase of 10,000 steps. Optimization was performed using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, ϵ = 1e-9). The final model, implemented in Python with TensorFlow [43], contained 694,658 parameters and was used for both the separation and baseline cases. All computations were performed on an NVIDIA RTX 5000 GPU.

For the tree-based model, peptide sequences were encoded by concatenating sequence-derived features with metadata. Sequence features included amino acid counts, dipeptide, and tripeptide compositions. Metadata comprised peptide mass, mass-to-charge ratio, length, charge, and one-hot-encoded enzyme labels. This process resulted in feature vectors of length 18,285. The dataset was split into training (90%) and validation (10%) subsets with a fixed random seed (42) for reproducibility. Hyperparameters for the separation-specific and baseline cases were optimized using Hyperopt [113]. The Python wrapper of the XGBoost library was employed for model training and predictions.

4.5.8 Data availability

MaxQuant results and supplementary movies have been deposited at Mendeley Data under https://data.mendeley.com/datasets/szrn5srhyw.

4.5.9 Code availability

Custom code used for the data analysis has been deposited at https://github.com/cox-labs/CCS.

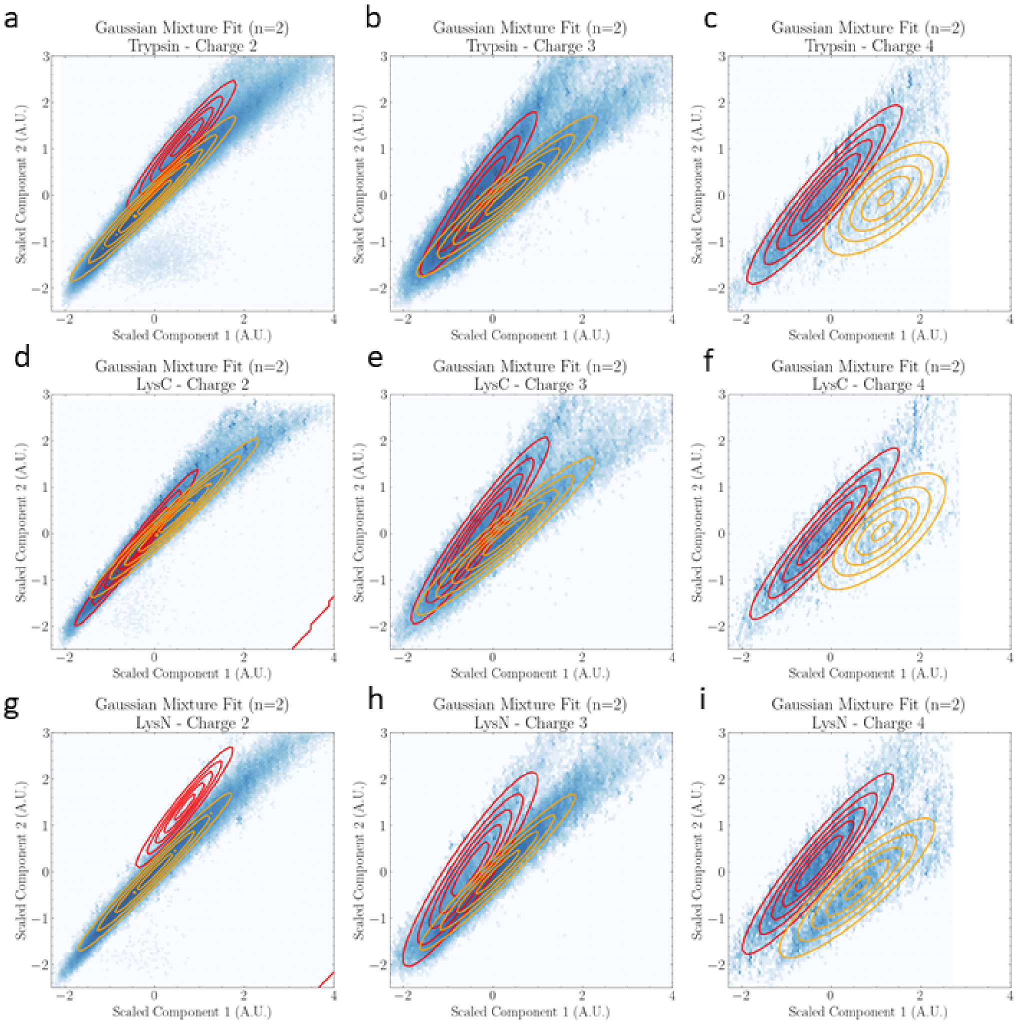
4.6 Acknowledgments

This work was supported by the German Ministry for Science and Education funding action CLINSPECT-M [FKZ 161L0214E].

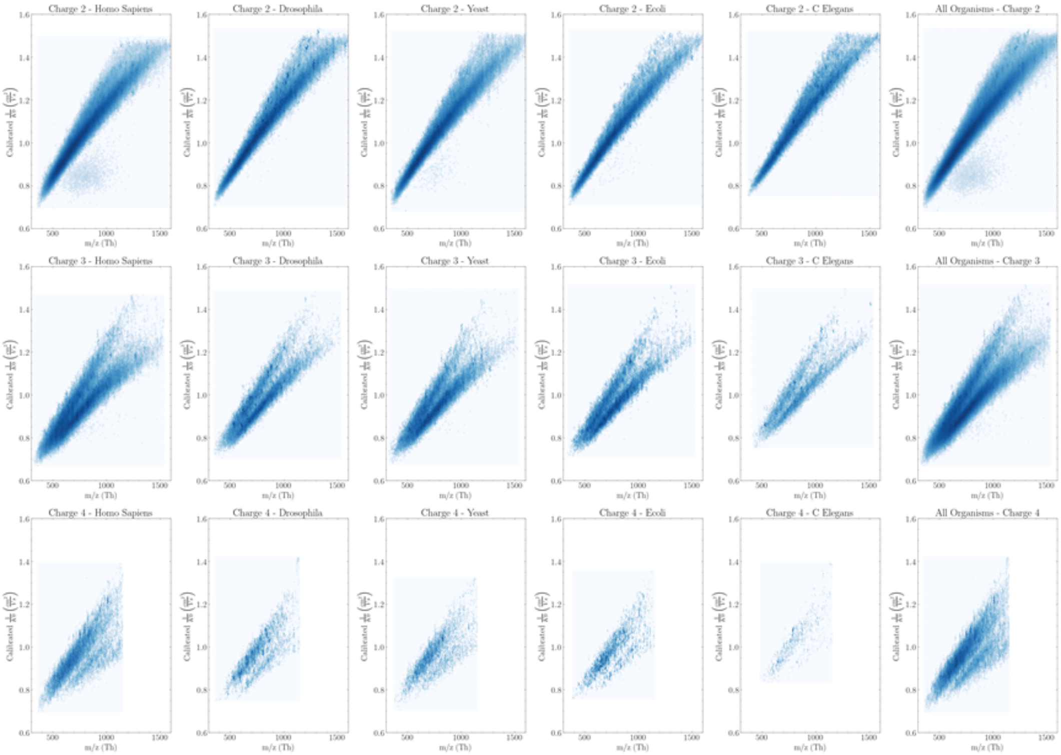
4.7 Author Contributions

J.R. did machine learning-based data analysis. D.S. performed the MD simulations, T.K. did the IMoS analysis, J.R., D.S. and T.K. did all other data analysis, J.R., D.S., C.W., H.G. and J.C wrote the manuscript, C.W., H.G. and J.C. directed the project. All authors read and approved the final manuscript.

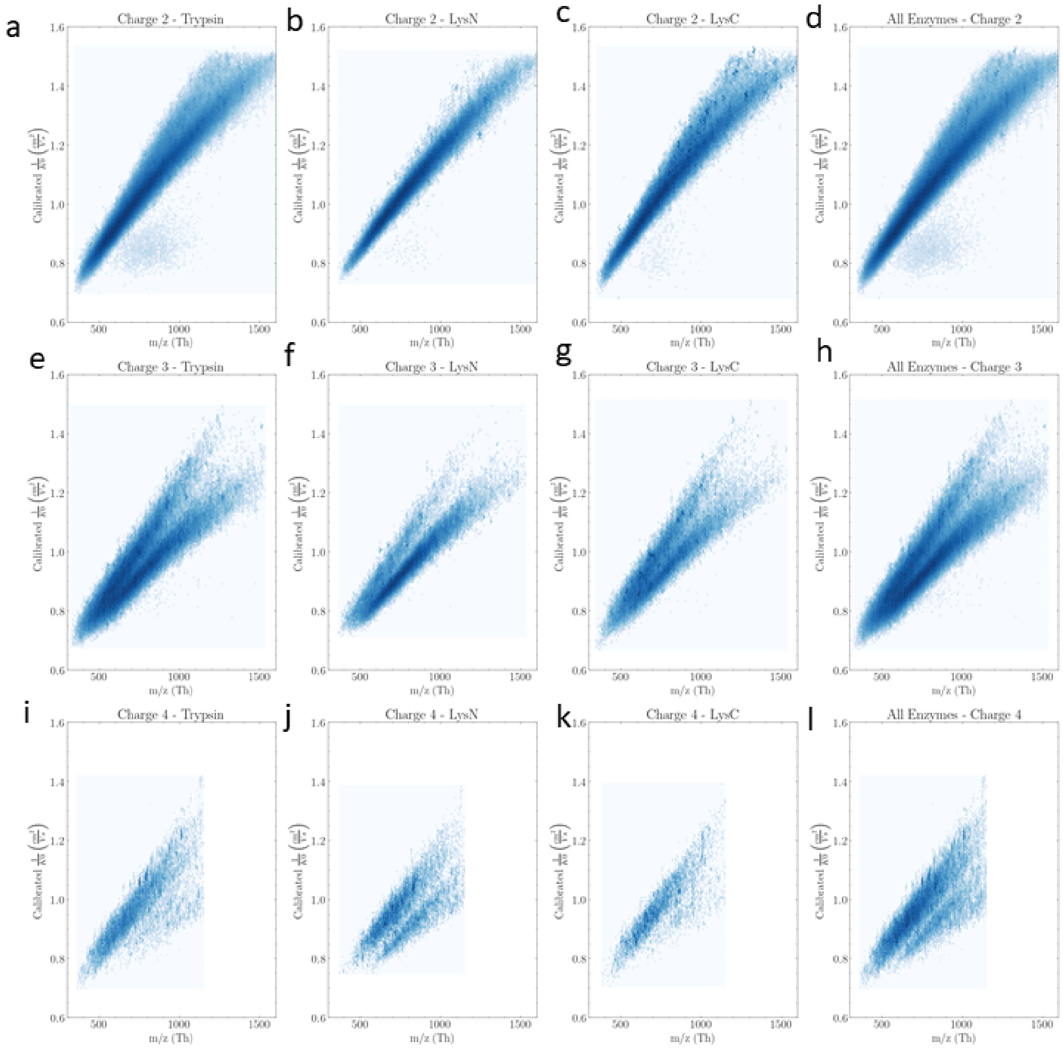
4.8 Supplementary Information



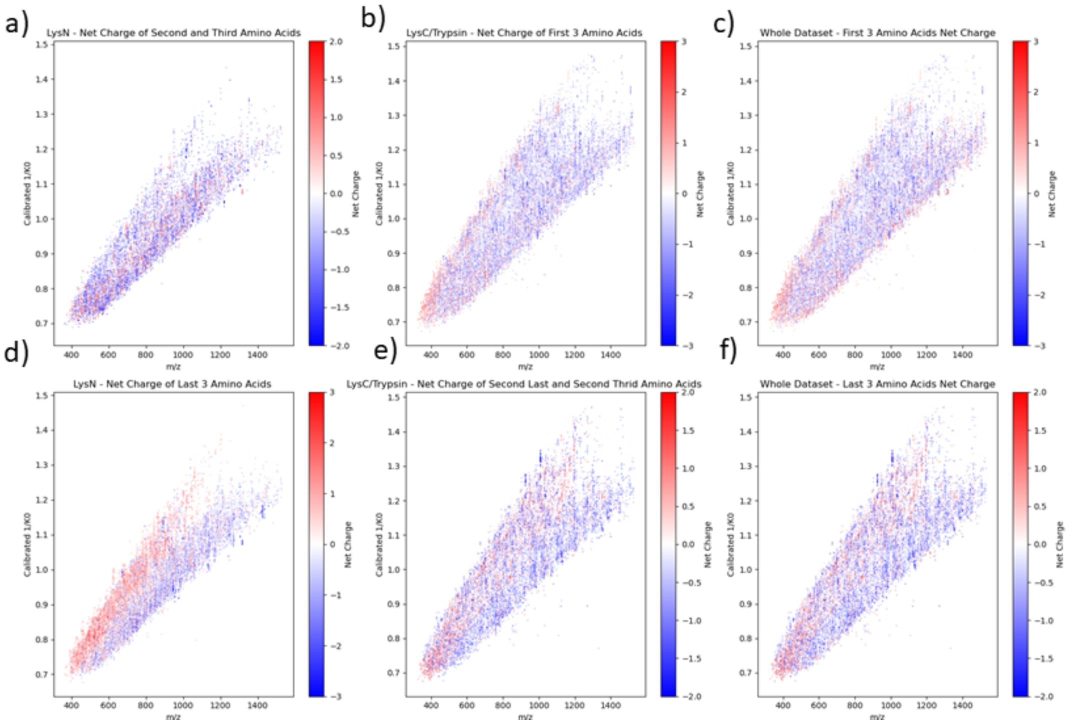
Supplementary Figure 4.1. Fit to the sum of two bivariate normal distributions per protease and charge state overlaid on the corresponding distribution. The distributions are normalized to have zero mean and unit variance.



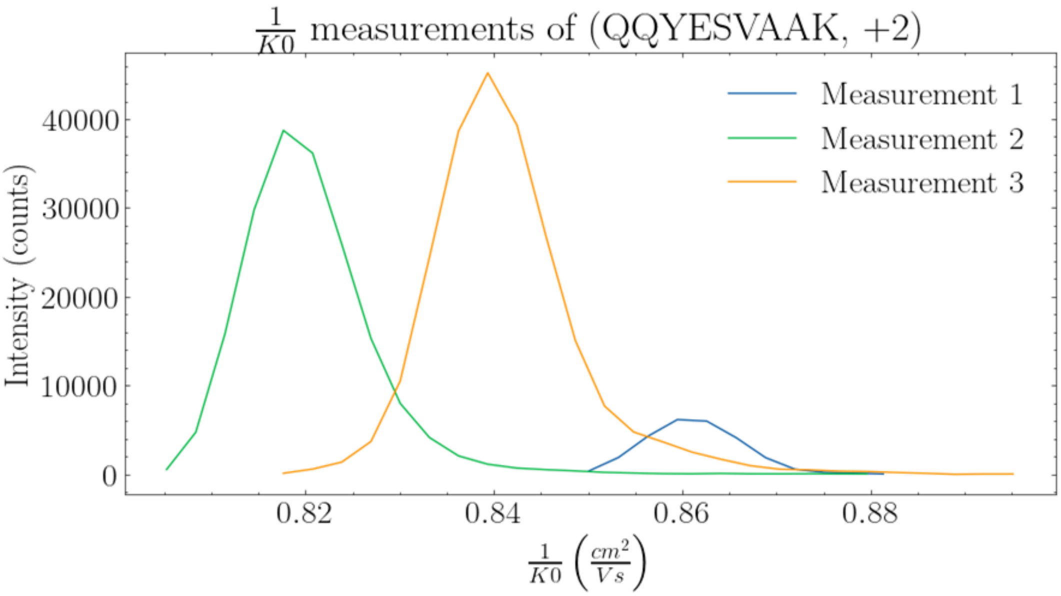
Supplementary Figure 4.2. Distribution of peptides in the space of reduced mobility calibrated across all runs versus mass-to-charge ratio for all charges and organisms.



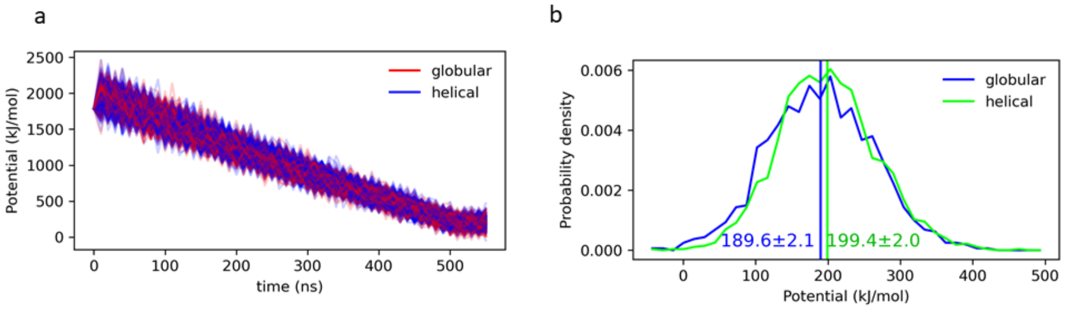
Supplementary Figure 4.3. Distribution of peptides in the space of reduced mobility calibrated across all runs versus mass-to-charge ratio for all charges and proteases.



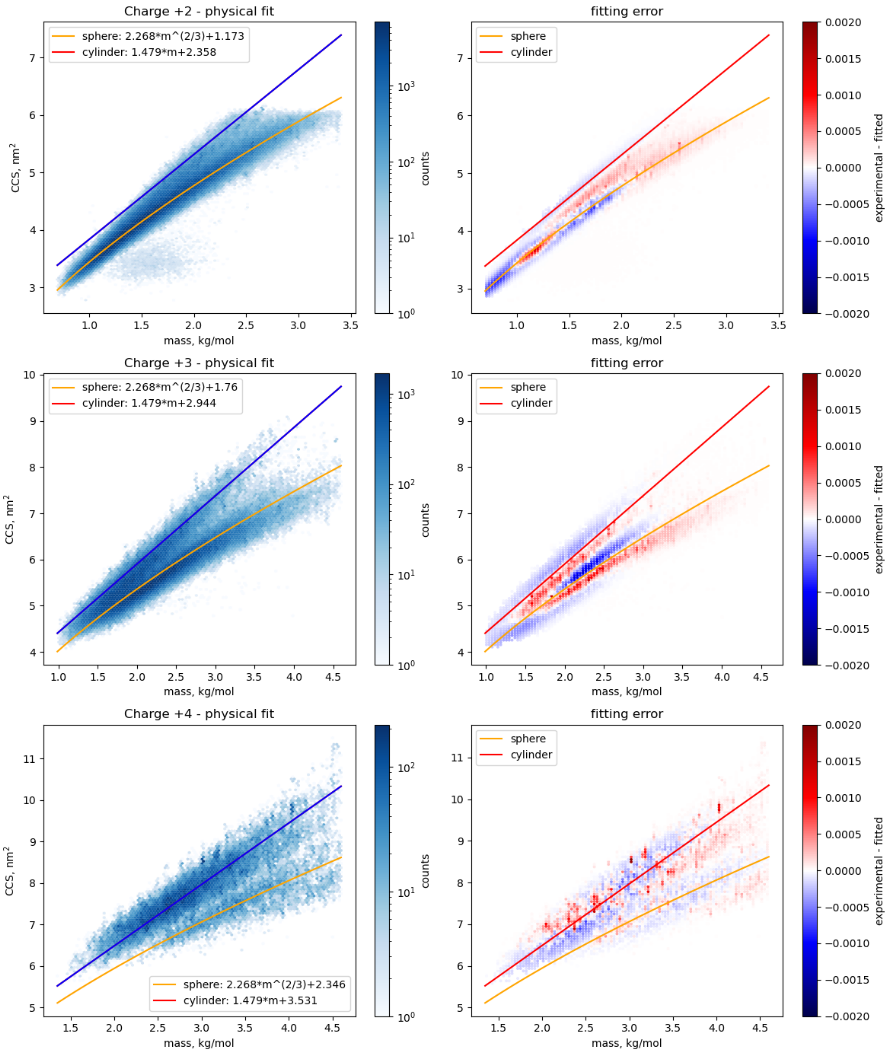
Supplementary Figure 4.4. Distribution of peptides in the space of reduced mobility calibrated across all runs versus mass-to-charge ratio for two enzymatic groups: LysC/Trypsin and LysN. Each distribution is colored by net charge in either the first three or the last three amino acids ignoring the charge associated to the protease. The distribution of the total dataset without enzymatic division is also shown on the right-most column.



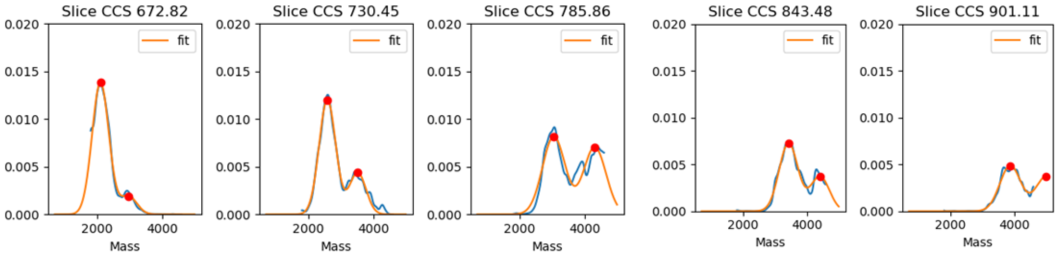
Supplementary Figure 4.5. Intensity profile on the reduced mobility dimension for a particular precursor (see title) that exhibits multiple peaks.



Supplementary Figure 4.6. Potential energy histogram of globular and helical conformations of the P1 peptide. Vertical lines and label indicate the mean potential energy and its standard error. The last 40 ns of the quenching simulations at constant temperature 305 K were used for this analysis. Grouping into globular (blue) and helical (green) was based on predicted CCS for the final conformation, structures above 850 Å^2 were considered helical, otherwise globular.



Supplementary Figure 4.7. Geometric fit (left column) and fitting error distribution (right column) for each charge state.



Supplementary Figure 4.8. Selected transversal slices with constant CCS over the mass dimension(blue line) for charge state four together with the fit to the sum of two one-dimensional gaussian distributions (orange line).

5

Back to Basics: Spectrum and Peptide Sequence are Sufficient for Top-tier Mass Spectrometry Proteomics Identification

This chapter presents a study focused on the development of a novel ML-based re-scoring model for DDA proteomics data. The work aims to improve PSM discrimination by leveraging features generated from MaxQuant outputs and applying advanced re-ranking strategies. The project represents a step toward more accurate and reproducible downstream analysis in shotgun proteomics pipelines, with a particular emphasis on enhancing post-search statistical modeling.

The outcomes of this project are available in a preprint hosted on *bioRxiv* [114], and the manuscript is currently under peer review at *Nature Methods*.

My contributions, made in collaboration with the team of the research group of *Computational Systems Biochemistry*, spanned multiple stages of development: Defining the project's scope with a clear focus on enhancing re-scoring techniques for DDA data, advising on the choice of public datasets used for training and benchmarking the model, Participating in the initial stages of codebase development and model prototyping, running *MaxQuant* to generate training data, as well as performing critical post-processing of the search results, contributing actively to the manuscript's writing and scientific framing.

The result of this collaboration is a robust framework that complements existing search engines by improving PSM validation while remaining compatible with standard proteomics workflows. The complete article, as it is online, is shown below.

Back to Basics: Spectrum and Peptide Sequence are Sufficient for Top-tier Mass Spectrometry Proteomics Identification

Maximilien Burq[1,#], Dejan Stepec[1,#], Juan Restrepo[2,#], Jure Zbontar[1], Shamil Urazbakhtin[2], Bryan Crampton[1], Shivani Tiwary[1], Rehan Chinoy[1], Melissa Miao[1], Jürgen Cox[2] and Peter Cimermancic1[1*]
[1]Tesorai Inc., Del Mar Heights Road Suite 284, San Diego, CA 92130, United States
[2]Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.
[#]These authors contributed equally to the publication.
[*]Correspondence: peter@tesorai.com

5.1 Abstract

The original mass spectrometry search engines used simple algorithms for peptide identification. Recent tools improved accuracy by adding several extra components such as fragment ion intensities or retention times prediction and training target-decoy classifiers on-the-fly, leading to sometimes inconsistent results.

Our study explores the impact of replacing those extra components with a deep-learning pretrained model that directly learns the complex relationship between the full spectra and associated peptide sequence, without using decoys. This simplified workflow has fewer parameters to tweak, making it easier to use and perform robustly on data from instruments and use-cases never seen during training.

Surprisingly, our approach consistently identifies more peptides than FragPipe, PEAKS, and Proteome Discoverer (12%, 9%, and 21% more, respectively, across a range of datasets). Tesorai Search is also fast – 250 immunopeptidomics searches in 45 minutes – and free for academics, available as a webserver at console.tesorai.com.

5.2 Introduction

Mass spectrometry (MS)-based proteomics typically involves breaking down complex protein mixtures into smaller peptides, which are then ionized and measured by a mass spectrometer. The instrument records the mass-to-charge ratio (m/z) of these intact peptides and subsequently fragments them to generate tandem mass spectra (MS2). Database search engines are the cornerstone computational tools used to interpret these complex spectra to determine the original peptide sequences. They work by comparing the experimentally acquired spectra against theoretical spectra generated from known protein sequences stored in databases. This process typically involves filtering potential peptide candidates based on precursor mass and then scoring the match between the experimental fragment ions and the predicted fragments for each candidate, resulting in peptide-spectrum matches (PSMs). The accuracy and comprehensiveness of protein identification are critically dependent on the performance of these search engines.

Original database search engines for mass spectrometry proteomics, such as Comet [115], MaxQuant (Andromeda) [105], [106], MSFragger [116], MS-GF+ [117], Proteome Discoverer [118] and Sequest [119], rely solely on spectral data — using precursor m/z for candidate selection and employing simple barcode matching with fixed canonical fragment ion intensities for scoring peptide-spectrum matches (PSMs). More recent search engines and rescoring tools such as MSBooster [120], MS2Rescore [121], PEAKS [122], Sage [123], Prosit/Oktoberfest [124, 125], Inferys[126] or Chimerys [127], have dramatically improved peptide identification rates with scoring innovations that incorporated dozens to hundreds of hand-crafted features based on prediction of canonical fragment ion intensities, differences between predicted and measured retention time, peptide lengths, and amino acid composition, followed by on-the-fly training of target-decoy classifiers such as Percolator[128] or PeptideProphet [129]. This added complexity increases the overfitting, which can in turn lead to underestimating the true false-discovery rate (FDR) in the generated search outputs[130].

We postulate that the full potential of proteomics identification can be achieved by returning to a spectrum-only approach — one that fully exploits the rich information in the entire spectrum alongside the peptide sequence. We propose that the limitations of the traditional methods arose not from the inadequacy of spectral data per se, but rather from incomplete utilization of the spectrum's information, insufficient incorporation of peptide sequence context, and reliance on scoring functions that fail to capture complex, non-linear associations between spectra and sequence. To test that hypothesis, we introduce an end-to-end deep-learning model designed to receive the complete spectrum and peptide sequence and directly output a calibrated score for a PSM. This model, scaled to capture chemical complexity and trained on heterogeneous datasets covering various enzymes, instruments, and fragmentation types, employs a novel loss function that avoids reliance on decoy training. This is especially important as models get larger and the risk that the model will learn entire proteomes increases. Our results demonstrate that a spectrum-based approach not only achieves high accuracy and sensitivity in peptide-spectrum matching but also surpasses the performance of both traditional and recent search engines. These results are in contrast to approaches in [131, 132, 133, 134] which share some of our ideas, but were unable to outperform modern tools for database search.

Achieving high performance without on-the-fly model retraining required several breakthroughs, both on the model architecture and training paradigm. Similarly to traditional search engines, we pre-compute theoretical fragment ion intensities for the candidate peptide sequence. The model, however, also has access to the full measured spectrum, including regions that are outside of those theoretical fragment m/z values. We propose a new modular architecture where the model propagates its own embedding vectors for each breakpoint, enabling it to implicitly learn features related to precursor fragmentation and fragment ion ionization probability. For training, we introduce a new paradigm where the model learns to discriminate correct PSMs from high-scoring incorrect ones. Thus, decoys are not used at any point during model training and are reserved for (FDR) estimation. This is critical to ensure that the model does not overfit to the decoy generation process, which would yield incorrect results.

Our secondary contribution is an extensive benchmarking of nine widely used database search engines, which we ran on diverse datasets, including tryptic digests, immunopeptidomics, and single-cell analyses. We find that Tesorai Search consistently identifies more unique peptides compared to established search engines like MaxQuant, FragPipe (+MSBooster), PEAKS, and Proteome Discoverer, achieving average increases of 12% to 68%. We hope that this benchmark can serve as the basis for accurate comparisons, spurring further innovation and enhancements of existing and new search engines.

Further experiments confirm the model's robustness, showing strong performance even on data not encountered during training, such as data produced with different instrument types (TOF) or using isobaric labeling (TMT). We verify through entrapment analyses that these increased identifications are achieved while maintaining accurate FDR control. Ablation studies confirm the model effectively utilizes information about peptide sequences and fragmentation spectra, including peaks coming outside of the canonical fragment ion series. Finally, we present a scalable, user-friendly cloud implementation capable of processing large datasets rapidly, making this advanced model readily accessible to the research community.

5.3 Results

5.3.1 Pretrained model paradigm

We hypothesized that training a single large peptide-spectrum matching algorithm can increase peptide identifications across use cases and reduce search engine complexity and data processing steps. To test this hypothesis, we propose a new algorithmic paradigm in which a deep-learning model outputs a score directly from a single tandem mass spectrum and candidate peptide sequence. This contrasts with existing ML-powered tools, which compute dozens to hundreds of metrics. In some cases, such as ion intensity or retention time prediction, each metric requires its own deep-learning model. These metrics then need to be combined into a score using a second-stage ML model – typically Percolator or PeptideProphet. Furthermore, to boost the performance, Percolator-based approaches incorporate information unrelated to peptide-spectrum matching, like charge and retention time error. We show that these additional features are unnecessary for accurate peptide-spectrum matching.

The Tesorai Search workflow (Fig. 5.1A) starts with user-provided raw mass spectrometry data (.d, .raw, or .mzml) and a protein sequence database (FASTA). Candidate PSMs are generated using Comet, MaxQuant, and MSFragger with permissive settings (100% FDR). Our pre-trained deep-learning model then rescores each spectrum-sequence pair, outputting a single score. These scores are subsequently used in a standard target-decoy competition strategy (see Methods) for accurate FDR control. The system outputs high-confidence PSMs and peptide identifications via a scalable cloud infrastructure.

Our approach simplifies processing by eliminating custom steps like deisotoping or recalibration, as these are implicitly learned by the model. The model architecture (Fig. 5.1B; Methods) has three components. A sequence encoder processes the modified peptide sequence (one-hot encoded), computes theoretical fragments, and combines them with a learned sequence embedding into a vector representation. A spectrum encoder generates a vector representation from the input centroided tandem mass spectrum (m/z and intensity lists). A joint encoder integrates the peptide and spectrum vectors, outputting a single numerical score indicating match quality.

We trained the model on a diverse dataset of ~289 million tandem mass spectra from 16,000 runs across 22 studies (Fig. 5.1C; Methods). Spectra were searched using MaxQuant, MS-GF+, and Comet to identify top-ranking (positive examples)

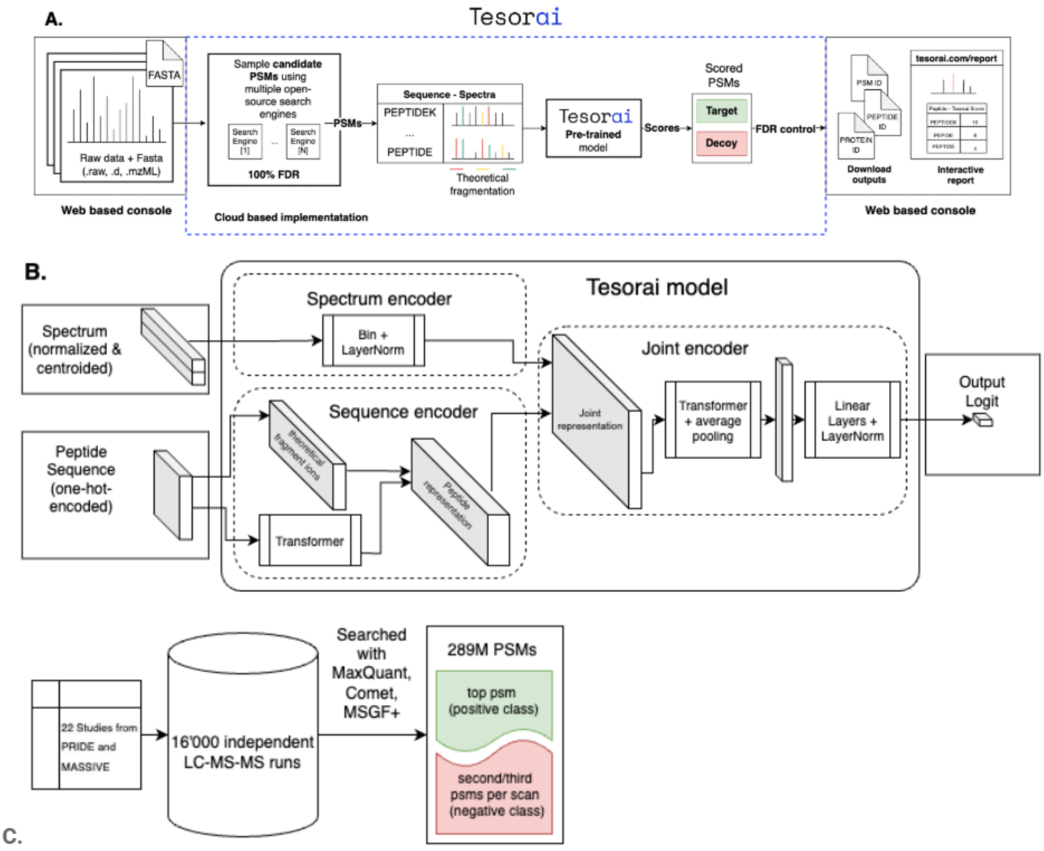
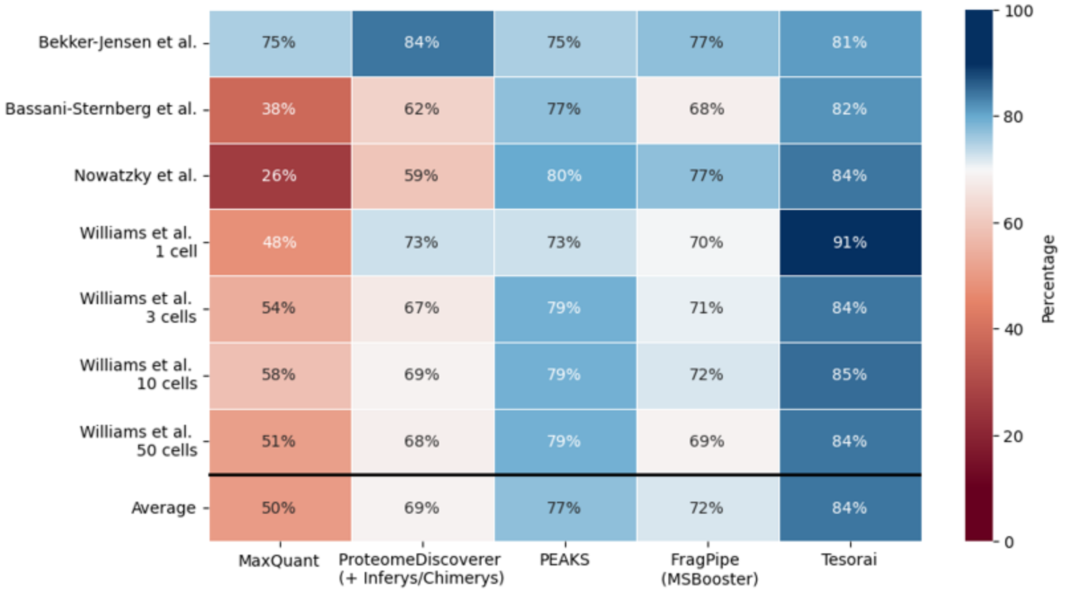


Figure 5.1. Overview of Tesorai Search. A. Rescoring pipeline. Illustrates the steps from raw data processing to rescoring with the pretrained model to the final list of identified peptides after FDR control. **B. Deep-learning-based rescoring model architecture.** Input data (MS2 spectrum and peptide sequence) is processed by the three key components, spectrum encoder, sequence encoder, and joint encoder, to generate a PSM score. **C. Training dataset generation and curation.** The training set consists of a pool of 22 studies and 16.000 LC-MS-MS runs processed with MaxQuant, Comet, and MS-GF+. The top PSM was taken as the positive class, and the second and third as the negative class

and lower-ranked (negative examples) peptide sequences. Importantly, the training occurs only once, does not utilize decoy sequences, and does not require retraining on the specific dataset of interest. Using real sequences in both positive and negative classes forces the model to learn genuine sequence-spectrum associations rather than memorize peptidomes.

5.3.2 Increased number of identifications across many common use cases

We benchmarked Tesorai Search against leading search engines: MaxQuant [105], FragPipe (with MSBooster [120]), Proteome Discoverer (using Chimerys/Inferys [129]— see Methods), and PEAKS [122] (Fig. 5.2). All tools, including Tesorai Search, were run with default settings on seven diverse datasets representing common proteomics use cases. To ensure fairness, some datasets were chosen from the original benchmark publications of the compared engines. Our open-source data-processing



	MaxQuant	ProteomeDiscoverer (+ Inferys/Chimerys)	PEAKS	FragPipe (MSBooster)	Tesorai
Bekker-Jensen et al.	75%	84%	75%	77%	81%
Bassani-Sternberg et al.	38%	62%	77%	68%	82%
Nowatzky et al.	26%	59%	80%	77%	84%
Williams et al. 1 cell	48%	73%	73%	70%	91%
Williams et al. 3 cells	54%	67%	79%	71%	84%
Williams et al. 10 cells	58%	69%	79%	72%	85%
Williams et al. 50 cells	51%	68%	79%	69%	84%
Average	50%	69%	77%	72%	84%

Figure 5.2. Unique peptide identifications (without modifications) at 1% FDR Numbers are shown relative to the union of all identified peptides across all search engines. **Bekker-Jensen** Deep fractionated tryptic HeLa sample [135]. **Bassani-Sternberg** Immunopeptidomics use-case [136]. **Nowatzky** Immunopeptidomics use-case; MSV000089312[137].**Williams** Single-cell samples (NanoPots) [138]. The bottom row shows the average across all samples.

code is available on GitHub[1].

We selected a deeply fractionated tryptic HeLa sample from Bekker-Jensen et al. [135], which was previously analyzed in both Prosit and MSBooster publications. This sample was searched against a human FASTA file obtained from the original Prosit publication. Subsequently, two immunopeptidomics samples were chosen. The first originated from Bassani-Sternberg et al. [136], where a sample designated Mel15 had been utilized in benchmarking Prosit and MSBooster. The second dataset was for Behçet's disease (Cavers et al.)[137]. Lastly, to assess our model's performance on samples with limited quantities, we incorporated a study encompassing single-cell and few-cell analysis[138], previously used to assess performance in the MSBooster publication.

On average, we identified 17% more (non-modified) peptides than FragPipe/MSBooster, 9% more than PEAKS, 21% more than Proteome Discoverer (+Chimerys/Inferys), and 68% more than MaxQuant. In Figure 5.2, we list the number of peptides identified by each search engine, normalized as a fraction of the total number of peptides identified across all search engines. A summary of identification counts and overlaps across search engines can be found in Supplementary Table 5.2. The data from runs across all search engines is available at Mendeley[139].

To further assess the robustness of the model, we evaluated its performance on datasets not included in our training set and distinct from our Orbitrap, label-free, bulk, non-modified training data (Supplementary Table 5.3). We compared Tesorai Search against FragPipe (MSBooster), selected for its strong performance and ease-of-

[1] https://github.com/tesorai/tesorai_search

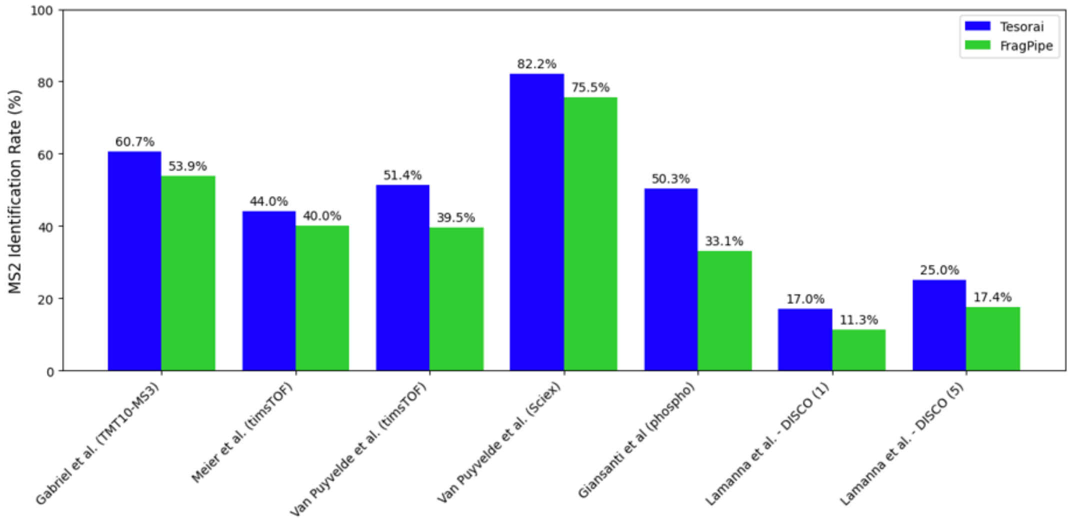


Figure 5.3. Scan (MS2) identification rate on data not seen in training. Comparison of MS2 scan identification rates at 1% FDR between Tesorai Search and FragPipe (MSBooster) across diverse proteomics datasets not seen during training: TMT10-MS3 (Gabriel et al.), TimsTOF instrument data (Meier et al.; Van Puyvelde et al.), Sciex data (Van Puyvelde et al.), phospho-enriched (Giasanti et al.), and single-cell DISCO (1 and 5 cells) (Lamanna et al.).

use in initial benchmarks. We compared the MS2 scan identification rates at a 1% FDR (Fig. 5.3). On a TMT-labeled dataset (Gabriel et al.) [140], we identified 13% more PSMs. On Time-of-Flight (TOF) instrument data, gains were +10% (Meier et al.) [141] and +30% (Van PuyVelde et al.) (Bruker timsTOF Pro), and +9% (Van PuyVelde et al.) (Sciex TripleTOF 6600+) [142]. Despite lacking specific PTM enrichment in training, Tesorai Search identified +51% PSMs on a phospho-enriched sample (Giasanti et al.). Furthermore, on challenging single-cell DISCO datasets (Lamanna et al. [143], 1 and 5 cell inputs), which differ significantly from bulk samples, we achieved 43-50% higher PSM identification rates. These results demonstrate that Tesorai Search generalizes effectively to diverse instruments, TMT labeling, PTM enrichment, and low-input single-cell data absent from its training.

5.3.3 Accurate false-discovery rate estimates

Increased identifications are meaningful only if the false-discovery rate (FDR) is correctly controlled, ensuring confidence in the results. We validated FDR control by measuring the false-discovery proportion (FDP), i.e. the true fraction of incorrect PSM or peptide matches, for user-defined FDR thresholds ranging from 0.1% to 10% (Fig. 5.4), ensuring FDP remained below the target FDR. Following [130, 144], we used entrapment analysis with the ISB18 dataset[145] (48 known target proteins: 18 synthesized, 30 contaminants) searched against a database combining these targets with the Ricinus communis (castor oil plant) proteome as an entrapment set chosen for low peptide overlap. PSM (Fig. 5.4A) and peptide (Fig. 5.4B) FDR were estimated using standard target-decoy competition. Given the high 668:1 entrapment-to-target peptide ratio, the FDP was computed as the number of identified entrapment sequences relative to non-entrapment sequences (known targets), using

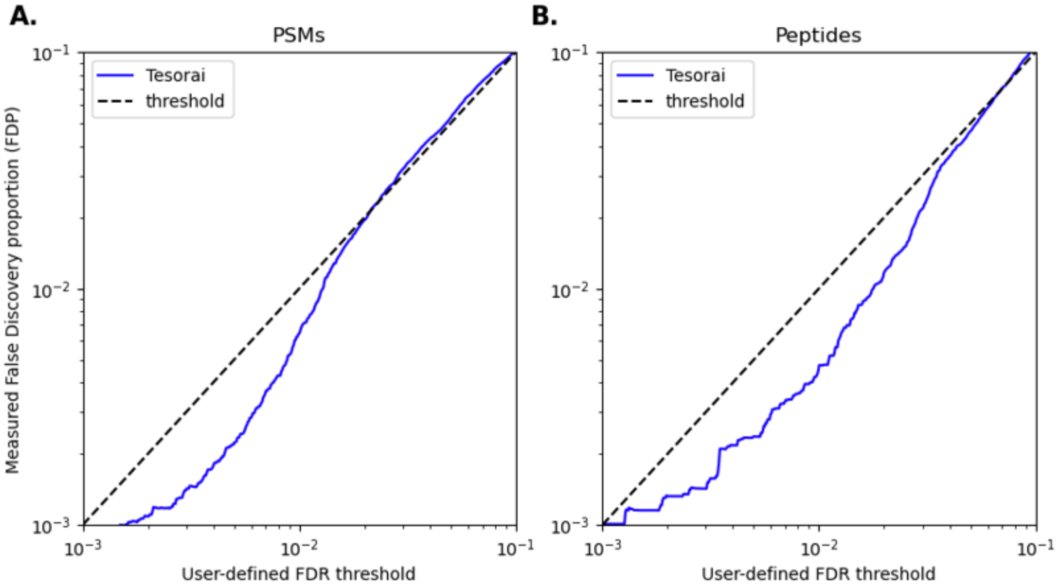


Figure 5.4. Measured false-discovery proportion (FDP) vs user-defined FDR threshold. A. FDP vs FDR at the PSM level of Tesorai Search (blue line), obtained through the entrapment method, compared to the user-defined false discovery threshold estimated using target-decoy competition in the region of interest, 0.1% to 10%. The dashed line represents the upper-bound case where the FDP is the same as the FDR.**B.** FDP vs FDR at the peptide level following the same evaluation approach.

the conservative "combined" formula [146]. Our results demonstrate that the fraction of erroneous PSMs and peptides identified by Tesorai Search is consistently lower than the user-defined FDR threshold, confirming accurate control.

5.3.4 Newly identified peptides

Having established accurate FDR control, we next characterized the novel peptide identifications by investigating the score distributions and whether new peptides are corroborated by other search engines. Figure 5.5A-B , shows score distributions for decoys (orange) and targets (blue) on the Mel15 sample [136]. Target scores follow a bimodal distribution (Fig. 5.5A-B, horizontal marginal distribution) where the first mode closely matches the distribution of decoy PSMs. This bimodal distribution, which is more pronounced than for the Andromeda score (Fig. 5.5A, vertical marginal) and MSFragger Hyperscore (Fig. 5.5B, vertical marginal), enables cleaner score separation and makes the resulting identifications less sensitive to the FDR cutoff threshold. Most new peptides are also identified by MaxQuant or MSFragger, albeit at a 5% FDR cutoff (Fig. 5.5C-D).

To further assess model reliability, we compared our peptide identifications against MaxQuant, Prosit, FragPipe, PEAKS, and Proteome Discoverer (Fig. 5.5E). For each tool, we categorized its identified peptides based on whether they were unique to that tool, shared with only one other tool, or shared with two or more other tools in the comparison set. For assessing our identifications, we did not count identifications that were only confirmed by FragPipe and MaxQuant, as these two tools are used by Tesorai Search to select initial PSMs for rescoring. Tesorai Search demonstrated the highest number of peptides confirmed by one or more other tools (70% of all identified peptides), compared to 37-65% for other tested tools.

5.3.5 Model interpretation

To understand the contributions of different input features and internal representations to the model's performance, we conducted an ablation study by corrupting model inputs and observing performance impacts (Fig. 5.6) On the PSM binary classification task used for model training, (validation set, ~41% positive, see Methods), the uncorrupted model achieved 99.5% accuracy. As a baseline representing complete information loss, shuffling both the peptide sequence input and the associated theoretical fragment m/z values reduced accuracy to 58.6%, equivalent to random guessing.

We assessed input importance by selectively randomizing features. Randomizing experimental spectrum intensities (keeping m/z fixed) substantially dropped accuracy to 77.6%, showing that the model uses intensity information beyond m/z matching. Filtering the spectrum to only expected canonical fragment ion regions (a, b, y, z) also significantly reduced accuracy (78.8%), indicating that the model utilizes information from non-canonical peaks or their absence.

Investigating the peptide-derived inputs revealed a strong dependence on the theoretical fragment masses. Randomizing these reduced the accuracy to 59.2% (near baseline). Shuffling only the separate one-hot peptide sequence encoding had a lesser impact (85.6% accuracy), highlighting the importance of theoretical fragment matching over abstract sequence encoding. Examining individual theoretical ion types revealed that randomizing y-ions caused the most degradation (72.2% accuracy), followed by b-ions (91.7%). Randomizing a-ions (99.2%) or z-ions (99.5%) had minimal effect, aligning with expected CID fragmentation importance. Full results are in Supplementary Table 5.4.

5.3.6 Cloud-based implementation allows for fast processing

To enhance processing speed and accessibility of proteomic analyses, we implemented a cloud-based interface for Tesorai Search. Users can perform database searches directly through a web browser without requiring local computational resources or software installations. We were able to re-process 250 immunopeptidomics samples from PXD019643 [147] in under 45 minutes, demonstrating the scalability of our implementation even on challenging workloads. This cloud-native approach not only simplifies workflows but also ensures that cutting-edge computational resources are available to researchers globally, facilitating rapid data processing and timely generation of results.

5.4 Discussion

This paper introduces Tesorai Search, an end-to-end deep learning model that fundamentally shifts the paradigm for peptide-spectrum matching in mass spectrometry proteomics. Our core hypothesis is that the full spectrum and peptide sequence

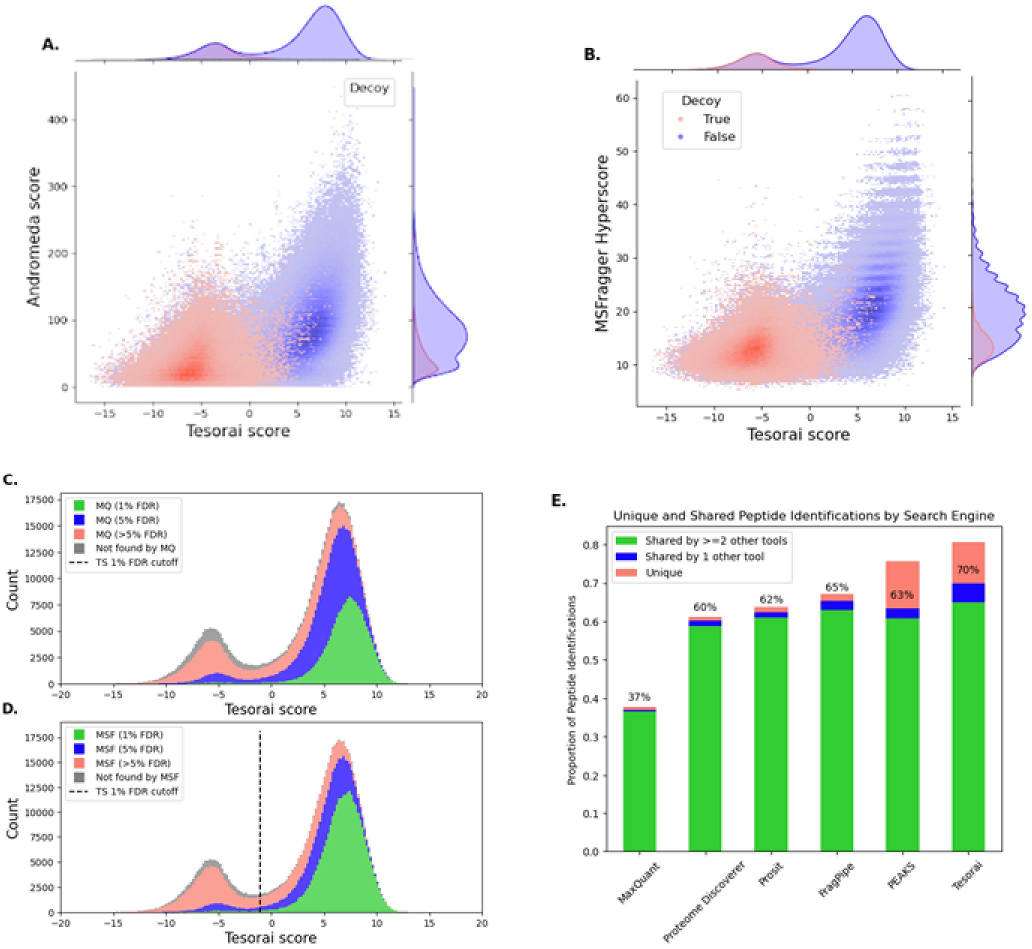


Figure 5.5. Newly identified peptides by origin using Bassani-Sternberg et al. (Mel15) dataset. A-B. Joint distribution of the Tesorai score against either the Andromeda score (A.) from MaxQuant or the MSFragger Hyperscore (B.). **C-D.** Breakdown of Tesorai Search identified PSMs by whether they were also identified by MaxQuant (C.) and MSFragger (D.) at various FDR thresholds. The dashed line represents the 1% FDR cutoff for the Tesorai Score. **E.** Breakdown of identified peptides from each tool, stratified by the number of other search engines that could confirm the identification. For Tesorai, we only included Proteome Discoverer and PEAKS in the tools used to confirm identification. We indicated the percentage of identified peptides that are shared with at least one other tool, as a percent of all peptides identified.

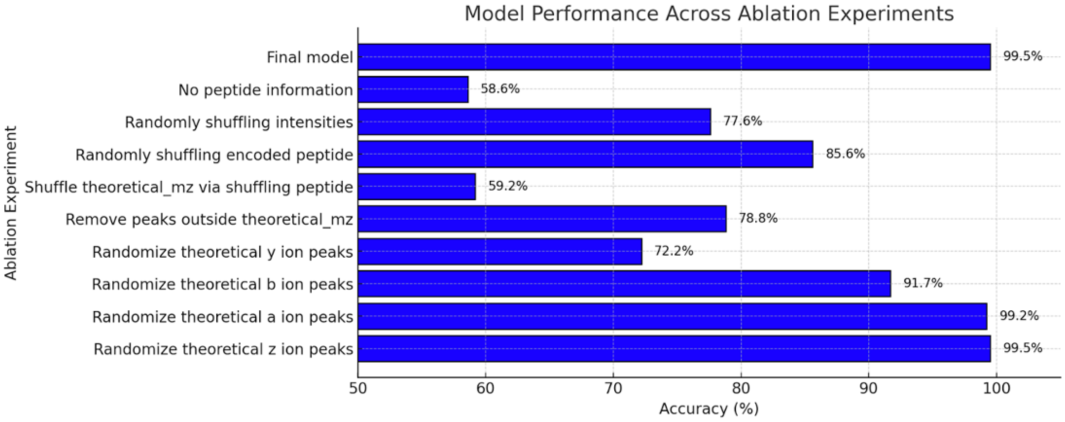


Figure 5.6. Effect of corrupting data inputs on model performance. Model accuracy (%) on the validation subset (41.4% positive, 58.6% negative) for the PSM binary classification task. Accuracy is shown for the full model (99.5%) and after selectively corrupting inputs via shuffling or randomization. No peptide information corresponds to shuffling all inputs.

contain sufficient information for top-tier peptide identification, challenging the need for intricate feature engineering or auxiliary predictive models (like retention time or fragment intensity prediction) used in recent tools. Our model directly learns complex, non-linear associations between complete MS/MS spectra and peptide sequences, demonstrating, through extensive benchmarking, that this approach surpasses both traditional and machine learning-enhanced search engines.

Existing machine learning-enhanced search engines often combine dozens of handcrafted or predicted features, frequently using post-processing classifiers like Percolator trained on-the-fly with target-decoy strategies. While improving upon traditional methods, this added complexity can introduce variability and potential inaccuracies, particularly concerning FDR estimation. The fundamental issue[130] is that Percolator uses decoys for two distinct purposes: training a machine learning model to differentiate between correct and incorrect identifications and estimating the FDR. A key conceptual advance of Tesorai Search is its ability to learn directly from real peptide-spectrum matches (PSMs) without relying on decoys for model training. Trained once on a large-scale dataset of ~289 million high-confidence, real PSMs, the model learns the genuine characteristics of correct identifications without potential biases introduced by artificial decoy sequences.

This results in a large increase in identification depth, with considerably more peptides identified than leading platforms like FragPipe, PEAKS, and Proteome Discoverer (12%, 9%, and 21%, respectively) across various use cases, including challenging immunopeptidomics and low-input samples. This enhanced sensitivity holds the potential for deeper biological insights [148]. For instance, in immunopeptidomics, identifying a broader repertoire of MHC-presented peptides is crucial for advancing cancer immunotherapy and vaccine design. The high proportion of identifications corroborated by other search engines supports their reliability. We also observed improved score separation compared to metrics like Andromeda, potentially due to better score calibration across different scans resulting from the classification-based training objective.

Furthermore, the model exhibits remarkable robustness and generalizability. Though trained exclusively on Orbitrap data, it performs well (identifying up to 50% more MS2 scans than FragPipe) on diverse sample types and data from instruments not included in the training set (TOF instruments, TMT-labeled samples). The end-to-end architecture implicitly learns necessary data processing steps, simplifying the workflow by obviating the need for explicit spectrum deisotoping, mass calibration, or charge correction. This inherent simplicity, combined with its implementation as a scalable, cloud-native platform, makes Tesorai Search readily accessible, facilitating rapid analysis of massive datasets and democratizing high-performance proteomics.

Several areas for future research emerge from this study. Though we focused exclusively on data-dependent acquisition workflows, the fundamental idea of using a single pretrained model for PSM rescoring is likely transferable to data-independent acquisition, promising simpler workflows, enhanced peptide identification, and improved FDR control in that setting as well. Although our method achieves state-of-the-art performance in the number of identified phosphopeptides, accurately localizing PTMs remains a challenge in the field. Furthermore, despite the advancements presented, overall scan identification rates can still be improved; combining our deep learning strategy with established techniques like match-between-runs and open window search approaches represents a promising direction for future research to maximize peptide recovery. Importantly, even with the strong concordance observed with existing tools, rigorous biological validation of uniquely identified peptides within relevant experimental contexts is essential to confirm their functional significance.

In conclusion, Tesorai Search demonstrates that a pre-trained deep learning model focused purely on the spectrum-sequence relationship can outperform other modern search engines. By offering enhanced sensitivity, robust FDR control, simplicity, and accessibility, this approach promises to accelerate discovery in fundamental biology and clinical proteomics.

5.5 Methods

5.5.1 Training dataset

For training, we identified 21 publicly available datasets (Supplementary Table 5.1), all processed with Orbitrap (Thermo Fisher Scientific) instruments. We first parsed the .raw files with ThermoRawFileparser to assemble the list of MS2 spectra. Each spectrum was represented as a tuple of two lists of floats, of equal lengths, representing the m/z value and the intensity of each (centroided) peak. No mass calibration or de-isotoping step was taken on the raw spectra.

In case the dataset was already processed with MaxQuant and an msms.txt file was publicly available, we used it directly, otherwise, we reprocessed the dataset with MaxQuant ourselves. Additionally, we processed some of the datasets with Comet and MS-GF+. For a given MS2 scan, MaxQuant provided the top-scoring peptide-spectrum match. This PSM was assigned to the positive class. Sometimes, second and third-best matches were also provided. These PSMs were assigned to the negative class. Altogether, this resulted in a total of 289 million PSMs in the training dataset – 38.4% positives and 61.6% negatives. We stored these spectra, the associated peptide sequences, and labels as TFRecords on disk to ensure low read times and high GPU utilization. To ensure that only high-quality PSMs were included in the training dataset, we filtered out those with an Andromeda score lower than 25, those with comet_score greater than 1e-3, and those with MS-GF+ SpecEValue greater than 1e-10.

We randomly split the peptide sequences based on mass into three splits: train, val, and test. Each peptide-spectrum match was then assigned to one of the buckets accordingly. Furthermore, we also reserved a set of PRIDE datasets where none of the PSMs were included in the train set. Finally, we created a list of benchmark datasets (see section "Processing of external data"), none of which were included in the training set, to measure the changes in identification rates.

5.5.2 Model architecture

The model consists of three components: a spectrum encoder, a sequence encoder, and a PSM score module. We used the PyTorch framework [44].

- **Spectrum processing** happens on the CPU. We take as input a float32 vector of m/z values and a float32 vector of intensities. Spectra with fewer than 10 peaks are discarded. For spectra with more than 1100 peaks, only the 1100 most intense peaks are kept. The intensities are then normalized so that they sum to 1. The spectra are then binned by 0.5 Th windows, and all intensities within a given bin are summed together. Peaks below 100 Th and above 2000 Th are discarded, resulting in a vector of 3800 bins. A LayerNorm[149] is then applied to the spectrum.
- **Sequence preprocessing** takes as input an int32 vector of one-hot-encoded amino-acid sequence of length s. It then computes theoretical fragment ions for a, b, y, z ions, two possible charge states (1 and 2), and optional water loss and ammonia loss. This results in 24 potential ions for each breakpoint in the peptide sequence. We used a total of 32 tokens to encode amino acids, where, in addition to the 20 canonical ones, we encoded the following 12 common modifications when computing theoretical fragment ions:
  - M(ox) for methionine oxidation, (mass=147.035405 Da)
  - Q(de) and N(de) for deamidation (mass += 0.9840 Da)
  - Z for n-term acetylation (mass = 42.0106 Da)
  - K(ac) for lysine acetylation (mass += 42.0106 Da)
  - E(py) for pyroglutamic acid (mass -= 18.01057 Da)
  - Q(py) for pyroglutamic acid (mass -= 17.02655 Da)
  - S(ph), T(ph) and Y(ph) for phosphorylation (mass += 79.966331 Da)
  - (tmt) and K(tmt) for isobaric labelling (variable mass shift depending on the --plex number)

  Furthermore, we trained our models with a fixed cysteine carbamidomethylation (+ 57.02146 Da).

5.5.3 Loss function and model training

The model was trained with a simple binary cross-entropy loss. We used AdamW[150] for gradient descent, torch.distributed for distributed data-parallel multi-GPU training, and torch.cuda.amp.autocast for automatic mixed-precision training on float16 weights, to speed up the training process.

The model was trained on a single node with 8 Nvidia Tesla V100-SXM2-16GB GPUs, with a batch size of 512 and 1M steps, which corresponds to two full epochs over the training dataset and took a total of 64 hours. We used the Pytorch framework.

5.5.4 Inference

For inference, the model takes as input a (sequence, spectrum) pair, and processes it in the same way as the training data. The spectrum is centroided and represented as a float32 vector of m/z values and a float32 vector of intensities. Spectra with fewer than 10 peaks are discarded. For spectra with more than 1100 peaks, only the 1100 most intense peaks are kept. The intensities are then normalized so that they sum to 1. The spectra are then binned by 0.5 Th windows, and all intensities within a given bin are summed together. Peaks below 100 Th and above 2000 Th are discarded, resulting in a vector of 3800 bins.

The sequence is one-hot encoded, and theoretical fragment ions are computed. It is then processed into the sequence encoder module, combined with the spectrum vector, and processed through the PSM score module.

We take the final number from the PSM score module as the score for the putative PSM, without passing it through the sigmoid function.

5.5.5 Search engine configurations

- **FragPipe configuration**

  FragPipe analysis was performed within the FragPipe v22.0 computational platform. FragPipe includes MSBooster, Percolator, and Philosopher [151] tools by default for the downstream processing of the MSFragger search results. Decoys were generated using Philosopher, within the FragPipe computational platform. The provided TMT10-MS3 FragPipe workflow setting was used for the Gabriel et al. dataset. FragPipe search was performed without contaminants. The analysis included variable modifications for methionine oxidation, N-terminal acetylation, and fixed cysteine carbamidomethylation. Dataset-specific parameters are described for each dataset separately. All other parameters were kept default across all the experiments.

- **Tesorai Search configuration**

  Searches using Tesorai were performed with the following default settings unless noted otherwise. The default enzyme selected was Trypsin/P (enzyme: Trypsin/P), configured for specific cleavage (enzyme_mode: SPECIFIC). By default, the search didn't include common contaminant sequences (include_contaminants: FALSE). Peptide identifications were filtered based on a default FDR threshold of 1% (fdr_threshold: 0.01). The minimum considered peptide length was set to 7 amino acids (min_peptide_length: 7). The maximum peptide length depended on the enzyme mode: for the default specific digestion, it was 63 amino acids (max_peptide_length: 63), while for unspecific digestion, the default maximum length was 25 amino acids. Default variable modifications are Methionine Oxidation and N-terminal Acetylation (variable_modifications: OXIDATION_M,ACETYLATION_NTERM). Cysteine Carbamidometylation is also enabled by default (static_modifications: CARBAMIDOMETHYL_C).

- **PEAKS configuration**

  PEAKS v12.5 was used with default settings. De novo search was disabled. Minimum and maximum peptide lengths were set to 7 and 50, respectively, except in immunopeptidomics runs, where we used 8 and 15.

- **Proteome Discoverer configuration.**

  Proteome Discoverer v3.0 was used, and Chimerys rescoring was enabled, except in immunopeptidomics samples where it is not supported and on which Inferys was used. Minimum and maximum peptide lengths were set to 7 and 50, respectively, except in immunopeptidomics runs, where we used 8 and 15.

5.5.6 Processing of external data.

- **Bekker-Jensen (HeLa) tryptic dataset**

  Following [119], the Bekker-Jensen et al. [135] multi-protease dataset was downloaded from the PRIDE repository with the identifier PXD004452. Files mapping to the identifier QE3_UPLC9_DBJ_SA_46fractions were selected for analysis. Prosit results were obtained at PXD010871, in the zipped folder Figure_2_3_5_Multiprotease_Dataset, under the directory: /trypsin/percolator_unzipped /prosit_target.psms. MaxQuant results were obtained at PXD004452, under the directory: SearchResults/msms.txt.

  The default settings were used for Tesorai search, except that the peptide length range was set to 7 - 50. All other settings were left to their default values. A human Swiss-Prot protein sequence database including annotated isoforms (downloaded May 3rd, 2024; 42421 protein sequences) was used for processing.

  FragPipe results were run by us using the settings described above. Enzymatic digestion was set to stricttrypsin, and the peptide length range was set to 7 - 50.

- **Bassani-Sternberg et al. (HLA I immunopeptides) dataset**

  Following Wilhelm et al. [152], the HLA Class I sample from patient Mel15 [136] was downloaded from the PRIDE repository with the identifier PXD004894. Only files mapping to HLA Class I from patient Mel15 were used in the analysis. The sequence database (HUMAN_2014) was provided within the zipped Search folder in PXD004894, which was used in all experiments.

  MaxQuant and Prosit results were obtained from PXD021398. MaxQuant results (msms.txt) were obtained from the zipped folder Figure_5_Mel15_MaxQuant1. Prosit results (prosit_target.peptides) were obtained from the zipped folder Figure_5_Mel15_MaxQuant100_and_Rescoring, under the directory Pride_100%/ rescoring_for_paper_2/percolator.

Tesorai search enzyme mode was set to UNSPECIFIC with the minimum peptide length set to 8 and maximum to 15, following [152].

FragPipe results were run by us using the settings described above. Enzymatic digestion was set to NONSPECIFIC. The minimum peptide length was set to 8 and the maximum to 15.

- **Nowatzky et al. (HLA I immunopeptides) dataset**

  The Nowatzky et al. dataset was downloaded from MSV000089312. Only files mapping to identifiers 190514_H_FreyaLC_AD_Nowatzky_HLA_test (KO and WT) were used in the analysis.

  Tesorai search enzyme mode was set to UNSPECIFIC with the minimum peptide length set to 7 and the maximum to 15.

  FragPipe results were run by us using the settings described above. Enzymatic digestion was set to NONSPECIFIC. The minimum peptide length was set to 7 and the maximum to 12.

  Search was performed with a combined FASTA of Human and a few select pathogen proteomes (Uniprotkb, accessed 2023_12_07).

- **Meier et al. (HeLa sample, timsTOF)**

  Human cervical cancer cell (HeLa) dataset used in [136] was downloaded from the PRIDE repository with the identifier PXD010012. The 200ng, 100ms experiment was reproduced with the raw files provided in the zipped folder HeLa_200ng_100ms_raw. MaxQuant results (msms.txt) were obtained in the same PRIDE repository in the zipped folder HeLa_200ng_100ms_txt.

  The default settings were used for Tesorai search. A reference human proteome (UP000005640) was downloaded from UniProt using one protein sequence per gene (March 26th, 2024; 20,590 protein sequences).

  FragPipe results were run by us using the settings described above. Enzymatic digestion was set to stricttrypsin and the peptide length range was set to 7 - 63.

- **Gabriel et al. (HeLa-Yeast, TMT)**

  Human Yeast Dilution TMT labeled dataset [140] was downloaded from the PRIDE repository with the identifier PXD030340. The HCD fragmentation on the Orbitrap mass analyzer (HCD/OT) experiment was reproduced. The 190416_FPTMT_MS3_HCDOT_R1 raw file was used in the analysis. MaxQuant (msms.txt) and MaxQuant with Andromeda (andromeda_target.pepides) results were obtained in the same PRIDE repository in the zipped folder HCD_OT_Without_rescoring. Prosit rescored results (prosit_target.peptides) were obtained in the same PRIDE repository in the zipped folder HCD_OT_With_rescoring.

  The default settings were used for Tesorai search, except the peptide length range was set to 7 to 50. A reference human proteome (UP000005640) was downloaded from UniProt using one protein sequence per gene (March 26th, 2024; 20,590 protein sequences). A reference Baker's yeast proteome (UP000002311) was downloaded from UniProt using one protein sequence per gene (March 28th, 2024; 6,060 protein sequences).

MSFragger results were run by us using the settings described above. Enzymatic digestion was set to stricttrypsin. The provided TMT10-MS3 FragPipe workflow setting was used and the peptide length range was set to 7 - 50.

- **Williams et al. (Single-cell NanoPOTS)**

  Following Yang et al. [120], the single-cell data from the nanoPOTS platform [138] was downloaded from the MassIVE repository under the identifier MSV000085230. We reproduced the experiment from using 1, 3, 10, and 50 cells from the MCF10A 30-minute experiment. We used the human Swiss-Prot protein sequence database provided at MSV000085230 for all the experiments (UniprotKB_homosapiens_Swiss_Prot_122916).

  The default settings were used for Tesorai search. FragPipe results were run by us using the settings described above. Enzymatic digestion was set to stricttrypsin and the peptide length range was set to 7 - 63.

- **G. Van Puyvelde et al. (BrukertimsTOF, Sciex)**

  Thermo Orbitrap QE HF-X, Bruker timsTOF Pro, and Sciex TripleTOF 6600+ [142] data was downloaded from the PRIDE repository with the identifier PXD028735. Raw files mapping to LFQ_[timsTOFPro_PASEF, TTOF6600]_DDA_Condition_A_Sample_[Alpha, Beta, Gamma]_[01-04]were selected for the analysis. Sciex .wiff files were peak-picked with MSConvert [153] and converted to mzML file format as our current platform does not natively support .wiff format. A reference database containing the Human, Yeast, and E.coli protein sequences was downloaded from UniProt in May 2024.

  The default Tesorai search settings were used, except the peptide length range was set to 7 - 50. FragPipe results were run by us using the settings described above. Enzymatic digestion was set to stricttrypsin and the peptide length range was set to 7 - 50.

- **H. Lamanna et al. (Single-cell DISCO)**

  Following [120], the single-cell data from the DISCO platform [138] was downloaded from the PRIDE repository under the identifier PXD019958. We reproduced the experiment from [120] using 1 and 5 cells Orbitrap Thermo Q-Exactive HF-X raw files, using all the 3 replicas for each experiment. We used the human Swiss-Prot protein sequence database provided at MSV000085230 for all the experiments (UniprotKB_homosapiens_Swiss_Prot_122916) that was also used in the Single-cell NanoPOTS experiment.

  The default Tesorai search settings were used.FragPipe results were run by us using the settings described above. Enzymatic digestion was set to stricttrypsin and the peptide length range was set to 7 - 63.

- **I. Giansanti et al. (phosphopeptidomics) dataset**

  The following six Orbitrap files were downloaded from the PRIDE repository under the identifier PXD001428 from Giansanti et al. [154]:

  - OR8_130622_TT_Trypsin_Ti-IMAC_Rep1_B1.raw

- OR8_130622_TT_Trypsin_Ti-IMAC_Rep1_B2.raw
- OR9_20130628_TT_Trypsin_Batch2_R1.raw
- OR9_20130628_TT_Trypsin_Batch2_R2.raw
- OR9_20130628_TT_Trypsin_Batch3_R1.raw
- OR9_20130628_TT_Trypsin_Batch3_R2.raw

The default settings were used for Tesorai search with an additional variable modifications enabled for phosphorylation modification of serine (S), threonine (T), and tyrosine (Y) residues (PHOSPHORYLATION_STY) and peptide length range between 7 and 50.

FragPipe was run with the default settings and phosphorylation STY variable modification enabled in the MSFragger search parameters. Peptide length range was between 7 and 50.

5.5.7 Cloud-based implementation

We enabled fast and scalable processing by running with high parallelism and leveraging on-demand, serverless Cloud resources. By default, each sample file is first run through a combination of FragPipe (MSFragger), MaxQuant and Comet at 100% FDR to generate a list of candidate PSMs for each MS2 scan. For immunopeptidomics runs, we disable MaxQuant and Comet to speed-up processing. This comes with almost no reduction in performance. Data collected on Sciex (.wiff) or Bruker (.d) instruments also is processed by MSFragger only.

Every file and algorithm is run on a separate Cloud machine, with optimized resources for that stage. The pre-trained model runs on T4 Nvidia GPUs and rescores the candidate PSMs.

These rescored PSMs are then filtered at 1% FDR, and assembled into peptide and protein lists. We use an open-source FFIA [155] quantification algorithm for all Orbitrap, Sciex and mzML datasets. For Bruker, we leverage the default quantification in FragPipe (IonQuant) [156].

The final results are surfaced in an intuitive and easy-to-use web interface, and full psm, peptide and protein tables are downloadable directly from the platform.

5.5.8 FDR estimation

We estimate FDR with the standard target-decoy approach. We use the default settings of each search engine to generate decoy sequences from the user-provided FASTA file. For MaxQuant and FragPipe, this is done by reversing the protein sequences prior to in-silico digestion. For Comet, we kept default settings: Comet generates decoys by reversing each target peptide sequence, keeping the N-terminal or C-terminal amino acid in place. We do not use protein-level information for additional filtering (sequential FDR).

5.5.9 Entrapment experiment to measure false-discovery proportion.

The raw data for the entrapment analysis was accessed on March 15, 2024 at https://regis-web.systemsbiology.net/PublicDatasets/18_Mix/Mix_7/LTQ/RAW_Data/. Following [146], we analyzed files 2-10 and excluded number 11. The Castor plant proteome was accessed from Uniprot on 2024-04-25. The list of ISB18 proteins was downloaded from https://regis-web.systemsbiology.net/PublicDatasets/database/18mix.fasta.

Following [146], we used strict tryptic enzymatic in-silico digestion (this is our default), and set missed cleavages to 0 (default 2). To reduce the probability of a contaminant randomly matching with a peptide from the castor plant, we set the minimum peptide length to 8. All other settings were kept default. All peptide mappings accounted for Isoleucine to Leucine substitution.

5.5.10 Data availability

All raw data analyzed in this study were already publicly available. All results run by ourselves, from Tesorai Search, MaxQuant, FragPipe, PEAKS, and Proteome Discoverer will be made available on Mendeley Data [139].

5.5.11 Code availability

The Tesorai platform, used to generate the main results, is available online at console.tesorai.com. The code used to process the results and generate figures and tables is available at https://github.com/tesorai/tesorai_search.

5.6 Author contributions

The Computational Systems Biochemistry Research Group, represented by J. R., S. U., and J. C., contributed essential proteomics expertise to the project. They played a key role in defining the project scope, which focused on developing a re-scoring model for data-dependent acquisition (DDA) data. The team provided guidance on the selection of datasets for both model development and benchmarking, and contributed to the early stages of codebase development and initial model prototyping. J. R. and S. U. assisted with running MaxQuant to generate the training dataset and performed post-processing of the results. Additionally, J. R., S. U., and J. C. were actively involved in writing the manuscript.

5.7 Supplementary Information

PRIDE Project	**Sample Count**	**Species**	**Sample Type**
PXD024364	65349	Homo sapiens (human)	Cell
PXD010154	53374	Homo sapiens (human)	Tissue
PXD021013	29284	Homo sapiens (human)	Synthetic
PXD003668	24565	Homo sapiens (human)	Cell
PXD013615	19756	Homo sapiens (human)	Cell

PRIDE Project	Sample Count	Species	Sample Type
PXD004732	16264	Homo sapiens (human)	Synthetic
PXD037285	16035	Homo sapiens (human)	Cell
PXD014877	13426	Multiple*	Not sure
PXD023119	10835	Homo sapiens (human)	Synthetic
PXD010595	10602	Homo sapiens (human)	Synthetic
PXD005353	6370	Homo sapiens (human)	Cell
PXD019483	6133	Homo sapiens (human)	Cell
PXD023120	5367	Homo sapiens (human)	Synthetic
PXD020079	3260	Homo sapiens (human)	Cell
PXD001608	2299	Homo sapiens (human)	Tissue
PXD004977	1505	Homo sapiens (human)	Cell
PXD000955	1347	Saccharomyces cerevisiae (baker's yeast)	Cell
PXD001695	1231	Saccharomyces cerevisiae (baker's yeast)	Cell
PXD008127	783	Homo sapiens (human)	Tissue
PXD020011	660	Homo sapiens (human)	Cell
PXD001865	434	Saccharomyces cerevisiae (baker's yeast)	Cell
PXD002452	33	Mus musculus (mouse)	Cell

Supplementary Table 5.1. Summary of PRIDE project data by species. This table lists the sample counts for different PRIDE projects along with their associated species.

	MaxQuant	Proteome Discoverer	FragPipe	Peaks	Tesorai	Union
bassani_sternberg	22166	35882	39410	44396	47393	57935
bekker_jensen	169476	190362	174795	170604	183147	224741
nowatzkv	3677	8333	10848	11255	11787	14107
williams/1_cells	1626	1964	1910	1892	2472	2761
williams/3_cells	2758	3435	3611	4048	4301	5114
williams/10_cells	4488	5382	5620	6130	6614	7766
williams/50_cells	7890	10482	10720	12285	13030	15472

Supplementary Table 5.2. Protein identifications across different tools and datasets.

Category	Dataset	FragPipe	Tesorai
TMT	Gabriel et al. (TMT10-MS3)	6,291	**6,748**
TOF instruments	Meier et al. (timsTOF)	67,155	**68,377**
	Van Puyvelde et al. (timsTOF)	68,065	**75,199**
	Van Puyvelde et al. (Sciex)	34,809	**40,886**
Phosphopeptidomics	Giansanti et al.	12,588	**14,387**
Single-cell(s)Lamanna et al. - DISCO	1 cell	5,713	**6,744**
	5 cells	9,598	**11,134**
Immunopeptidomics	Sarkizova et al.	11,841	**12,063**

Supplementary Table 5.3. FragPipe vs Tesorai

Ablation experiment	**Accuracy**
Final model	**99.5**
No peptide information (Shuffle peptide and theoretical_mz)	58.6
Randomly shuffling intensities	77.6
Randomly shuffling encoded peptide	85.6
Shuffle theoretical_mz via shuffling peptide	59.2
Remove peaks in real spectrum outside of theoretical_mz fragment ions (a,b,y,z)	78.8
Randomize theoretical y ion peaks	72.2
Randomize theoretical b ion peaks	91.7
Randomize theoretical a ion peaks	99.2
Randomize theoretical z ion peaks	99.5

Supplementary Table 5.4. Ablation experiment

Part III

Conclusions

The overarching goal of this thesis was to explore how machine learning can be synergistically integrated with physics-based modeling to enhance computational workflows in mass spectrometry-based proteomics. Two central research questions guided this work:

1. Can the incorporation of peptide structural information, particularly gas-phase conformations, improve peptide identification in mass spectrometry-based proteomics?
2. Is it possible to develop a general-purpose machine learning model that re-scores PSMs using only spectral and sequence information, while maintaining both sensitivity and reliability across datasets?

To address the first question, we investigated the phenomenon of bimodal CCS distributions observed in ion mobility spectrometry. Through a combination of molecular dynamics simulations, theoretical modeling, and machine learning-based CCS prediction, we demonstrated that certain peptides exhibit distinct and stable conformations in the gas phase. By explicitly modeling this bimodality, we showed that peptide matching accuracy improves, particularly in cases where conformational ambiguity leads to misidentification. This result emphasizes the role of subtle physicochemical properties in shaping mass spectrometric observables and suggests that structure-aware representations can significantly reduce search space ambiguity.

To answer the second question, we developed a discriminative machine learning model capable of re-scoring PSMs using only peptide sequences and their associated fragmentation spectra. This model, trained on multiple datasets and designed to require minimal feature engineering, achieved state-of-the-art identification performance while preserving accurate FDR control. The approach relies solely on spectrum-sequence pairs, demonstrating that high-confidence identifications are possible without auxiliary metadata or complex preprocessing pipelines. The success of this approach underscores the potential of deep learning models to capture essential biochemical patterns directly from raw data.

Together, these two contributions support the broader hypothesis that integrating data-driven models with physically grounded representations enhances the accuracy, robustness, and interpretability of proteomic identification pipelines. In both projects, the interplay between machine learning and domain-specific physical insights proved critical in addressing core challenges in proteomics: ambiguity in identification, limitations of existing scoring functions, and generalizability across instruments and sample types.

The findings presented in this thesis open several promising directions for future research. The demonstrated value of CCS bimodality modeling encourages further exploration of peptide structure ensembles, particularly with respect to charge localization, sequence motifs, and post-translational modifications. Future work may include incorporating predicted three-dimensional structures, such as those obtained from AlphaFold or molecular dynamics simulations, into identification scoring.

As MS continues to integrate orthogonal modalities-such as IMS and RT-future models could jointly embed multi-modal data for holistic identification. Furthermore, building on the success of pretrained sequence models in natural language processing, there is considerable potential in adapting large-scale protein language models to learn task-specific representations for PSM scoring, de novo sequencing, or MS/MS intensity prediction.

Finally, given the scale and complexity of modern proteomics datasets, it is essential to develop computational tools that are not only performant but also reproducible and scalable. The cloud-based implementation explored in the second project could be extended into modular, API-accessible pipelines that support large-scale analyses and collaborative or clinical research environments.

In summary, this thesis demonstrates that the integration of machine learning and physical modeling holds significant promise for advancing mass spectrometry-based proteomics. The methodologies developed here provide a foundation for future innovations that aim to increase identification depth, confidence, and interpretability in complex biological samples.

Acronyms

BPTT backpropagation through time

CCS Collision Cross Section

CID Collision-Induced Dissociation

Da Dalton

DDA Data-Dependent Acquisition

DIA Data-Independent Acquisition

DL Deep Learning

ESI Electrospray Ionization

FDR False Discovery Rate

GRU Gated Recurrent Units

HPLC High-Performance Liquid Chromatography

IMS Ion Mobility Spectrometry

LSTM Long Short-Term Memory

MD Molecular Dynamics

ML Machine Learning

MS Mass Spectrometry

PSM Peptide-Spectrum Match

RNN Recurrent Neural Network

RT Retention Time

List of Figures

Bibliography

[1] Ruedi Aebersold and Matthias Mann. "Mass spectrometry-based proteomics". In: *Nature* 422.6928 (Mar. 2003), pp. 198–207. ISSN: 1476-4687. DOI: 10.1038/nature01511. URL: http://dx.doi.org/10.1038/nature01511.

[2] Timothy K. Toby, Luca Fornelli, and Neil L. Kelleher. "Progress in Top-Down Proteomics and the Analysis of Proteoforms". In: *Annual Review of Analytical Chemistry* 9.Volume 9, 2016 (2016), pp. 499–519. ISSN: 1936-1335. DOI: https://doi.org/10.1146/annurev-anchem-071015-041550. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-anchem-071015-041550.

[3] George R. Waller and Otis C. Dermer. "Editorial". In: *Mass Spectrometry Reviews* 1.1 (Mar. 1982), pp. 1–2. ISSN: 1098-2787. DOI: 10.1002/mas.1280010102. URL: http://dx.doi.org/10.1002/mas.1280010102.

[4] Jürgen H. Gross. *Mass Spectrometry: A Textbook*. Springer Berlin Heidelberg, 2011. ISBN: 9783642107115. DOI: 10.1007/978-3-642-10711-5. URL: http://dx.doi.org/10.1007/978-3-642-10711-5.

[5] Arjun Ravikumar, Adrian Arrieta, and Chang C Liu. "An orthogonal DNA replication system in yeast". In: *Nature Chemical Biology* 10.3 (Feb. 2014), pp. 175–177. ISSN: 1552-4469. DOI: 10.1038/nchembio.1439. URL: http://dx.doi.org/10.1038/nchembio.1439.

[6] Hanno Steen and Matthias Mann. "The abc's (and xyz's) of peptide sequencing". In: *Nature Reviews Molecular Cell Biology* 5.9 (Sept. 2004), pp. 699–711. ISSN: 1471-0080. DOI: 10.1038/nrm1468. URL: http://dx.doi.org/10.1038/nrm1468.

[7] Lloyd R. Snyder, Joseph J. Kirkland, and John W. Dolan. *Introduction to Modern Liquid Chromatography*. Wiley, Nov. 2009. ISBN: 9780470508183. DOI: 10.1002/9780470508183. URL: http://dx.doi.org/10.1002/9780470508183.

[8] Lane C. Sander and Stephen A. Wise. "Synthesis and characterization of polymeric C18 stationary phases for liquid chromatography". In: *Analytical Chemistry* 56.3 (Mar. 1984), pp. 504–510. ISSN: 1520-6882. DOI: 10.1021/ac00267a047. URL: http://dx.doi.org/10.1021/ac00267a047.

[9] U. D. Neue. "HPLC Columns, Theory, Technology, and Practice". In: *Instrumentation Science & Technology* 26.4 (1998), pp. 439–440. DOI: 10.1080/10739149808001913. URL: https://doi.org/10.1080/10739149808001913.

[10] Lars Konermann et al. "Unraveling the mechanism of electrospray ionization". In: *Analytical Chemistry* 85.1 (Jan. 2013), pp. 2–9.

[11] Paul Kebarle and Udo H Verkerk. "Electrospray: from ions in solution to ions in the gas phase, what we know now". In: *Mass Spectrometry Reviews* 28.6 (Nov. 2009), pp. 898–917.

[12] John B. Fenn et al. “Electrospray Ionization for Mass Spectrometry of Large Biomolecules”. In: *Science* 246.4926 (Oct. 1989), pp. 64–71. ISSN: 1095-9203. DOI: 10.1126/science.2675315. URL: http://dx.doi.org/10.1126/science.2675315.

[13] Miriam Sannomiya et al. “Application of liquid chromatography/electrospray ionization tandem mass spectrometry to the analysis of polyphenolic compounds from an infusion of Byrsonima crassa Niedenzu”. In: *Rapid Communications in Mass Spectrometry* 19.16 (2005), pp. 2244–2250.

[14] M. Karas, U. Bahr, and T. Dülcks. “Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine”. In: *Fresenius Journal of Analytical Chemistry* 366.6-7 (Mar. 2000), pp. 669–676.

[15] Francesco Lanucara et al. “The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics”. In: *Nature Chemistry* 6.4 (Mar. 2014), pp. 281–294. ISSN: 1755-4349. DOI: 10.1038/nchem.1889. URL: http://dx.doi.org/10.1038/nchem.1889.

[16] Valérie Gabelica and Erik Marklund. “Fundamentals of ion mobility spectrometry”. In: *Current Opinion in Chemical Biology* 42 (Feb. 2018), pp. 51–59. ISSN: 1367-5931. DOI: 10.1016/j.cbpa.2017.10.022. URL: http://dx.doi.org/10.1016/j.cbpa.2017.10.022.

[17] Brandon T. Ruotolo et al. “Ion Mobility–Mass Spectrometry Reveals Long-Lived, Unfolded Intermediates in the Dissociation of Protein Complexes”. In: *Angewandte Chemie International Edition* 46.42 (Oct. 2007), pp. 8001–8004. ISSN: 1521-3773. DOI: 10.1002/anie.200702161. URL: http://dx.doi.org/10.1002/anie.200702161.

[18] Jody C. May, Caleb B. Morris, and John A. McLean. “Ion Mobility Collision Cross Section Compendium”. In: *Analytical Chemistry* 89.2 (Dec. 2016), pp. 1032–1044. ISSN: 1520-6882. DOI: 10.1021/acs.analchem.6b04905. URL: http://dx.doi.org/10.1021/acs.analchem.6b04905.

[19] Edward A Mason and Homer W Schamp. “Mobility of gaseous lons in weak electric fields”. In: *Annals of Physics* 4.3 (July 1958), pp. 233–270. ISSN: 0003-4916. DOI: 10.1016/0003-4916(58)90049-6. URL: http://dx.doi.org/10.1016/0003-4916(58)90049-6.

[20] Leonardo Collado-Torres et al. “Reproducible RNA-seq analysis using recount2”. In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 319–321. ISSN: 1546-1696. DOI: 10.1038/nbt.3838. URL: http://dx.doi.org/10.1038/nbt.3838.

[21] Ludovic C. Gillet et al. “Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis”. In: *Molecular and Cellular Proteomics* 11.6 (June 2012), O111.016717. ISSN: 1535-9476. DOI: 10.1074/mcp.o111.016717. URL: http://dx.doi.org/10.1074/mcp.O111.016717.

[22] Reta Birhanu Kitata, Jhih-Ci Yang, and Yu-Ju Chen. “Advances in data-independent acquisition mass spectrometry towards comprehensive digital proteome landscape”. In: *Mass Spectrometry Reviews* 42.6 (May 2022), pp. 2324–2348. ISSN: 1098-2787. DOI: 10.1002/mas.21781. URL: http://dx.doi.org/10.1002/mas.21781.

[23] Christina Ludwig et al. "Data-independent acquisition-based <scp>SWATH</scp> - <scp>MS</scp> for quantitative proteomics: a tutorial". In: *Molecular Systems Biology* 14.8 (Aug. 2018). ISSN: 1744-4292. DOI: 10.15252/msb.20178126. URL: http://dx.doi.org/10.15252/msb.20178126.

[24] Sebastien Gallien, Elodie Duriez, and Bruno Domon. "Selected reaction monitoring applied to proteomics". In: *Journal of Mass Spectrometry* 46.3 (Mar. 2011), pp. 298–312. ISSN: 1096-9888. DOI: 10.1002/jms.1895. URL: http://dx.doi.org/10.1002/jms.1895.

[25] Jürgen Cox and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification". In: *Nature Biotechnology* 26.12 (Nov. 2008), pp. 1367–1372. ISSN: 1546-1696. DOI: 10.1038/nbt.1511. URL: http://dx.doi.org/10.1038/nbt.1511.

[26] Jürgen Cox et al. "Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment". In: *Journal of Proteome Research* 10.4 (Apr. 2011), pp. 1794–1805. ISSN: 1535-3907. DOI: 10.1021/pr101065j. URL: http://dx.doi.org/10.1021/pr101065j.

[27] Andy T Kong et al. "MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics". In: *Nature Methods* 14.5 (Apr. 2017), pp. 513–520. ISSN: 1548-7105. DOI: 10.1038/nmeth.4256. URL: http://dx.doi.org/10.1038/nmeth.4256.

[28] Benjamin C. Orsburn. "Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics". In: *Proteomes* 9.1 (Mar. 2021), p. 15. ISSN: 2227-7382. DOI: 10.3390/proteomes9010015. URL: http://dx.doi.org/10.3390/proteomes9010015.

[29] Vadim Demichev et al. "DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput". In: *Nature Methods* 17.1 (Nov. 2019), pp. 41–44. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0638-x. URL: http://dx.doi.org/10.1038/s41592-019-0638-x.

[30] Roland Bruderer et al. "Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues". In: *Molecular and Cellular Proteomics* 14.5 (May 2015), pp. 1400–1410. ISSN: 1535-9476. DOI: 10.1074/mcp.m114.044305. URL: http://dx.doi.org/10.1074/mcp.M114.044305.

[31] Pavel Sinitcyn et al. "MaxDIA enables library-based and library-free data-independent acquisition proteomics". In: *Nature Biotechnology* 39.12 (July 2021), pp. 1563–1573. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00968-7. URL: http://dx.doi.org/10.1038/s41587-021-00968-7.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". en. In: *Nature* 521.7553 (May 2015), pp. 436–444.

[33] Y Lecun et al. "Gradient-based learning applied to document recognition". In: *Proc. IEEE Inst. Electr. Electron. Eng.* 86.11 (1998), pp. 2278–2324.

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet classification with deep convolutional neural networks". en. In: *Commun. ACM* 60.6 (May 2017), pp. 84–90.

[35] Bo Wen et al. "Deep learning in proteomics". en. In: *Proteomics* 20.21-22 (Nov. 2020), e1900335.

[36] Shivani Tiwary et al. "High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis". en. In: *Nat. Methods* 16.6 (June 2019), pp. 519–525.

[37] Siegfried Gessulat et al. "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning". en. In: *Nat. Methods* 16.6 (June 2019), pp. 509–518.

[38] Ngoc Hieu Tran et al. "De novo peptide sequencing by deep learning". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.31 (Aug. 2017), pp. 8247–8252.

[39] Ngoc Hieu Tran et al. "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry". en. In: *Nat. Methods* 16.1 (Jan. 2019), pp. 63–66.

[40] J Elman. "Finding structure in time". In: *Cogn. Sci.* 14.2 (June 1990), pp. 179–211.

[41] Singh, Harinder and Singh, Sandeep and Singh Raghava, Gajendra Pal. *Peptide secondary structure prediction using evolutionary information*. bioRxiv. 2019. URL: %7Bhttps://www.biorxiv.org/content/10.1101/558791v1%7D.

[42] Jaspreet Singh et al. "SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning". en. In: *Bioinformatics* 37.20 (Oct. 2021), pp. 3464–3472.

[43] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. arXiv preprint arXiv:1603.04467. 2016. URL: https://arxiv.org/abs/1603.04467.

[44] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv preprint arXiv:1912.01703. 2019. URL: https://arxiv.org/abs/1912.01703.

[45] S Hochreiter and J Schmidhuber. "Long short-term memory". en. In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780.

[46] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. arXiv preprint arXiv:1406.1078. 2014. URL: https://arxiv.org/abs/1406.1078.

[47] Paul J. Werbos. "Backpropagation Through Time: What It Does and How to Do It". In: *Proceedings of the IEEE* 78.10 (1990), pp. 1550–1560. DOI: 10.1109/5.58337.

[48] Sangtae Kim et al. "SMSNet: Integrated de novo peptide sequencing and database search for single-cell proteomics". In: *Nature Methods* 15.7 (2018), pp. 515–518. DOI: 10.1038/s41592-018-0260-3. URL: https://www.nature.com/articles/s41592-018-0260-3.

[49] Yarin Gal and Zoubin Ghahramani. “A theoretically grounded application of dropout in recurrent neural networks”. In: *bioarxiv* (2015). eprint: 1512.05287 (stat.ML).

[50] Ashish Vaswani et al. “Attention is all you need”. In: *bioarxiv* (2017). eprint: 1706.03762 (cs.CL).

[51] Tom B Brown et al. *Language Models are Few-Shot Learners*. 2020. eprint: 2005.14165 (cs.CL).

[52] OpenAI et al. “GPT-4 Technical Report”. In: *bioarxiv* (2023). eprint: 2303.08774 (cs.CL).

[53] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0.

[54] Diederik P Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv preprint arXiv:1412.6980* (2014). URL: https://arxiv.org/abs/1412.6980.

[55] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958.

[56] Paulius Micikevicius et al. “Mixed Precision Training”. In: *arXiv preprint arXiv:1710.03740* (2018).

[57] Tri Dao et al. “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness”. In: *Advances in Neural Information Processing Systems*. 2022.

[58] Jeff Rasley et al. “DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters”. In: *arXiv preprint arXiv:2007.03029* (2020).

[59] Rui Qiao et al. “Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices”. en. In: *Nat. Mach. Intell.* 3.5 (Mar. 2021), pp. 420–425.

[60] Chunwei Ma et al. “Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning”. In: *Analytical Chemistry* 90.18 (2018), pp. 10881–10888.

[61] Pier-Luc Plante et al. “Predicting ion mobility collision cross-sections using a deep neural network: DeepCCS”. en. In: *Anal. Chem.* 91.8 (Apr. 2019), pp. 5191–5199.

[62] Ayano Nakai-Kasai et al. “Leveraging pretrained deep protein language model to predict peptide collision cross section”. en. In: *Commun. Chem.* 8.1 (May 2025), p. 137.

[63] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. Apr. 2019.

[64] Nadav Brandes et al. “ProteinBERT: a universal deep-learning model of protein sequence and function”. en. In: *Bioinformatics* 38.8 (Apr. 2022), pp. 2102–2110.

[65] Fangping Wan, Daphne Kontogiorgos-Heintz, and Cesar de la Fuente-Nunez. “Deep generative models for peptide design”. en. In: *Digit. Discov.* 1.3 (June 2022), pp. 195–208.

[66] Emre Sevgen et al. “ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design”. In: *bioRxiv* (Jan. 2023).

[67] Jael Sanyanda Wekesa and Michael Kimwele. “A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment”. en. In: *Front. Genet.* 14 (July 2023), p. 1199087.

[68] Jenna L Ballard et al. “Deep learning-based approaches for multi-omics data integration and analysis”. en. In: *BioData Min.* 17.1 (Oct. 2024), p. 38.

[69] Juan Restrepo et al. “Bimodal peptide collision cross section distribution reflects two stable conformations in the gas phase”. In: *bioRxiv* (2025). DOI: 10.1101/2025.05.19.654929. URL: https://www.biorxiv.org/content/10.1101/2025.05.19.654929v1.

[70] A.B. Kanu et al. “Ion mobility-mass spectrometry”. en. In: *Journal of Mass Spectrometry* 43 (2008). Preprint at, pp. 1–22. DOI: 10.1002/jms.1383.

[71] R. Cumeras et al. “Review on Ion Mobility Spectrometry. Part 2: hyphenated methods and effects of experimental parameters”. en. In: *Analyst* 140 (2015), pp. 1391–1410.

[72] J.C. May and J.A. McLean. “Ion mobility-mass spectrometry: Time-dispersive instrumentation”. en. In: *Analytical Chemistry* 87 (2015). 1422–1436 Preprint at. DOI: 10.1021/ac504720m. URL: https://doi.org/10.1021/ac504720m.

[73] V. Gabelica and E. Marklund. “Fundamentals of ion mobility spectrometry”. en. In: *Current Opinion in Chemical Biology* 42 (2018). Preprint at. DOI: 10.1016/j.cbpa.2017.10.022.

[74] D.A. Wolters, M.P. Washburn, and J.R. Yates. “An automated multidimensional protein identification technology for shotgun proteomics”. en. In: *Anal Chem* 73 (2001), pp. 5683–5690.

[75] Y. Zhang et al. *Protein analysis by shotgun/bottom-up proteomics*. en. Chemical Reviews Preprint at. 2013. DOI: 10.1021/cr3003533.

[76] R. Aebersold and M. Mann. “Mass-spectrometric exploration of proteome structure and function”. en. In: *Nature* 537 (2016), pp. 347–355.

[77] P. Sinitcyn, J.D. Rudolph, and J. Cox. “Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data”. no. In: *Annu Rev Biomed Data Sci* 1 (2018), pp. 207–234.

[78] A. Michalski, J. Cox, and M. Mann. “More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS”. en. In: *J Proteome Res* 10 (2011), pp. 1785–1793.

[79] S.J. Valentine. “Toward plasma proteome profiling with ion mobility-mass spectrometry”. en. In: *J Proteome Res* (2006). DOI: 10.1021/pr060232i..

[80] E.S. Baker. “An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies”. en. In: *J Proteome Res* (2010). DOI: 10.1021/pr900888b..

[81] S.J. Geromanos et al. “Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples”. en. In: *Anal Bioanal Chem* (2012). DOI: 10.1007/s00216-012-6197-y..

[82] D. Helm. “Ion Mobility Tandem Mass Spectrometry Enhances Performance of Bottom-up Proteomics”. su. In: *Molecular Cellular Proteomics* (2014). DOI: 10.1074/mcp.M114.041038..

[83] S.J. Valentine. “Using ion mobility data to improve peptide identification: Intrinsic amino acid size parameters”. en. In: *J Proteome Res* 10 (2011).

[84] A.E. Counterman and D.E. Clemmer. “Large anhydrous polyalanine ions: Evidence for extended helices and onset of a more compact state”. en. In: *J Am Chem Soc* 123 (2001).

[85] J.A. Silveira. “From solution to the gas phase: Stepwise dehydration and kinetic trapping of substance p reveals the origin of peptide conformations”. en. In: *J Am Chem Soc* 135 (2013).

[86] M. Dole. “Molecular beams of macroions”. af. In: *J Chem Phys* 49 (1968).

[87] C. Eldrid. “Gas Phase Stability of Protein Ions in a Cyclic Ion Mobility Spectrometry Traveling Wave Device”. en. In: *Anal Chem* 91 (2019).

[88] S. Nguyen and J.B. Fenn. “Gas-phase ions of solute species from charged droplets of solutions”. en. In: *Proc Natl Acad Sci U S A* 104 (2007).

[89] A.E. Counterman and D.E. Clemmer. “Large anhydrous polyalanine ions: Evidence for extended helices and onset of a more compact state”. en. In: *J Am Chem Soc* 123 (2001).

[90] L.W. Zilch et al. “Folding and Unfolding of Helix-Turn-Helix Motifs in the Gas Phase”. en. In: *J Am Soc Mass Spectrom* 18 (2007).

[91] B.S. Kinnear, M.R. Hartings, and M.F. Jarrold. “Helix unfolding in unsolvated peptides”. en. In: *J Am Chem Soc* 123 (2001).

[92] M. Kohtani et al. “Extreme stability of an unsolvated -helix”. en. In: *J Am Chem Soc* 126 (2004).

[93] Y. Wei, W. Nadler, and U.H.E. Hansmann. “On the helix-coil transition in alanine based polypeptides in gas phase”. en. In: *Journal of Chemical Physics* 126 (2007).

[94] K. Breuker and F.W. McLafferty. “Stepwise evolution of protein native structure with electrospray into the gas phase, 10-12 to 102 s”. en. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 105. Preprint at. 2008. DOI: 10.1073/pnas.0807005105.

[95] T. Meyer et al. “Proteins in the gas phase”. en. In: *Wiley Interdiscip Rev Comput Mol Sci* 3 (2013).

[96] C.A. Browne, H.P.J. Bennett, and S. Solomon. “The isolation of peptides by high-performance liquid chromatography using predicted elution positions”. en. In: *Anal Biochem* 124 (1982).

[97] S.J. Valentine, A.E. Counterman, and D.E. Clemmer. “A database of 660 peptide ion cross sections: use of intrinsic size parameters for bona fide predictions of cross sections”. en. In: *J Am Soc Mass Spectrom* 10 (1999).

[98] C.H. Chang. “Sequence-Specific Model for Predicting Peptide Collision Cross Section Values in Proteomic Ion Mobility Spectrometry”. en. In: *J Proteome Res* 20 (2021).

[99] R. Devreese. *Collisional cross-section prediction for multiconformational peptide ions with IM2Deep*. en. bioRxiv 2025.02.18.638865 (2025. DOI: 10.1101/2025.02.18.638865..

[100] F. Meier. “Deep learning the collisional cross sections of the peptide universe from a million experimental values”. en. In: *Nat Commun* 12 (2021).

[101] S.A. Ewing et al. “An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method”. en. In: *J Am Soc Mass Spectrom* 28 (2017).

[102] C. Larriba-Andaluz and C.J. Hogan. “Collision cross section calculations for polyatomic ions considering rotating diatomic/linear gas molecules”. en. In: *Journal of Chemical Physics* 141 (2014).

[103] J. Jumper. “Highly accurate protein structure prediction with AlphaFold”. en. In: *Nature* (2021). DOI: 10.1038/s41586-021-03819-2..

[104] S. Tiwary. “High quality MS/MS spectrum prediction for data-dependent and -independent acquisition data analysis”. en. In: *Nat Methods* (2019).

[105] J. Cox and M. Mann. “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification”. en. In: *Nat Biotechnol* 26 (2008), pp. 1367–1372.

[106] J. Cox. “Andromeda: a peptide search engine integrated into the MaxQuant environment”. en. In: *J Proteome Res* 10 (2011), pp. 1794–1805.

[107] Florian Meier et al. “Online parallel accumulation-serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer”. en. In: *Mol. Cell. Proteomics* 17.12 (Dec. 2018), pp. 2534–2545.

[108] H.P. Erickson. “Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy”. en. In: *Biological Procedures Online* 11 (2009). Preprint at. DOI: 10.1007/s12575-009-9008-x. URL: https://doi.org/10.1007/s12575-009-9008-x.

[109] G. Bussi, D. Donadio, and M. Parrinello. “Canonical sampling through velocity rescaling”. en. In: *Journal of Chemical Physics* 126 (2007).

[110] J. Huang. “CHARMM36m: An improved force field for folded and intrinsically disordered proteins”. en. In: *Nat Methods* 14 (2016).

[111] S. Wang, K. Hou, and H. Heinz. “Accurate and Compatible Force Fields for Molecular Oxygen, Nitrogen, and Hydrogen to Simulate Gases, Electrolytes, and Heterogeneous Interfaces”. en. In: *J Chem Theory Comput* 17 (2021).

[112] T. Chen and C. Guestrin. “XGBoost : Reliable Large-scale Tree Boosting System”. en. In: *ArXiv* (2016). DOI: 10.1145/2939672.2939785..

[113] James Bergstra et al. “Hyperopt: a Python library for model selection and hyperparameter optimization”. In: *Comput. Sci. Discov.* 8.1 (July 2015), p. 014008.

[114] Maximilien Burq et al. “Back to basics: Spectrum and peptide sequence are sufficient for top-tier mass spectrometry proteomics identification”. In: *bioRxiv* (2024). DOI: 10.1101/2024.08.19.606805. URL: https://www.biorxiv.org/content/10.1101/2024.08.19.606805v2.full.

[115] Jimmy K. Eng, Tahmina A. Jahan, and Michael R. Hoopmann. “Comet: An open-source <scp>MS</scp>/<scp>MS</scp> sequence database search tool”. In: *PROTEOMICS* 13.1 (Dec. 2012), pp. 22–24. ISSN: 1615-9861. DOI: 10.1002/pmic.201200439. URL: http://dx.doi.org/10.1002/pmic.201200439.

[116] Andy T Kong et al. “MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics”. In: *Nature Methods* 14.5 (Apr. 2017), pp. 513–520. ISSN: 1548-7105. DOI: 10.1038/nmeth.4256. URL: http://dx.doi.org/10.1038/nmeth.4256.

[117] Sangtae Kim and Pavel A. Pevzner. “MS-GF+ makes progress towards a universal database search tool for proteomics”. In: *Nature Communications* 5.1 (Oct. 2014). ISSN: 2041-1723. DOI: 10.1038/ncomms6277. URL: http://dx.doi.org/10.1038/ncomms6277.

[118] Benjamin C. Orsburn. “Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics”. In: *Proteomes* 9.1 (Mar. 2021), p. 15. ISSN: 2227-7382. DOI: 10.3390/proteomes9010015. URL: http://dx.doi.org/10.3390/proteomes9010015.

[119] Jimmy K. Eng, Ashley L. McCormack, and John R. Yates. “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database”. In: *Journal of the American Society for Mass Spectrometry* 5.11 (Nov. 1994), pp. 976–989. ISSN: 1044-0305. DOI: 10.1016/1044-0305(94)80016-2. URL: http://dx.doi.org/10.1016/1044-0305(94)80016-2.

[120] Kevin L. Yang et al. “MSBooster: improving peptide identification rates using deep learning-based features”. In: *Nature Communications* 14.1 (July 2023). ISSN: 2041-1723. DOI: 10.1038/s41467-023-40129-9. URL: http://dx.doi.org/10.1038/s41467-023-40129-9.

[121] Louise M. Buur et al. “MS2Rescore 3.0 Is a Modular, Flexible, and User-Friendly Platform to Boost Peptide Identifications, as Showcased with MS Amanda 3.0”. In: *Journal of Proteome Research* 23.8 (Mar. 2024), pp. 3200–3207. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.3c00785. URL: http://dx.doi.org/10.1021/acs.jproteome.3c00785.

[122] Bin Ma et al. “PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry”. In: *Rapid Communications in Mass Spectrometry* 17.20 (Sept. 2003), pp. 2337–2342. ISSN: 1097-0231. DOI: 10.1002/rcm.1196. URL: http://dx.doi.org/10.1002/rcm.1196.

[123] Michael R. Lazear. “Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale”. In: *Journal of Proteome Research* 22.11 (Oct. 2023), pp. 3652–3659. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.3c00486. URL: http://dx.doi.org/10.1021/acs.jproteome.3c00486.

[124] Siegfried Gessulat et al. "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning". In: *Nature Methods* 16.6 (May 2019), pp. 509–518. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0426-7. URL: http://dx.doi.org/10.1038/s41592-019-0426-7.

[125] Mario Picciani et al. "Oktoberfest: Open-source spectral library generation and rescoring pipeline based on Prosit". In: *PROTEOMICS* 24.8 (Sept. 2023). ISSN: 1615-9861. DOI: 10.1002/pmic.202300112. URL: http://dx.doi.org/10.1002/pmic.202300112.

[126] Daniel P. Zolg et al. "INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results". In: *Rapid Communications in Mass Spectrometry* 39.S1 (June 2021). ISSN: 1097-0231. DOI: 10.1002/rcm.9128. URL: http://dx.doi.org/10.1002/rcm.9128.

[127] Martin Frejno et al. "Unifying the analysis of bottom-up proteomics data with CHIMERYS". In: *Nature Methods* 22.5 (Apr. 2025), pp. 1017–1027. ISSN: 1548-7105. DOI: 10.1038/s41592-025-02663-w. URL: http://dx.doi.org/10.1038/s41592-025-02663-w.

[128] Lukas Käll et al. "Semi-supervised learning for peptide identification from shotgun proteomics datasets". In: *Nature Methods* 4.11 (Oct. 2007), pp. 923–925. ISSN: 1548-7105. DOI: 10.1038/nmeth1113. URL: http://dx.doi.org/10.1038/nmeth1113.

[129] Kelvin Ma, Olga Vitek, and Alexey I Nesvizhskii. "A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet". In: *BMC Bioinformatics* 13.S16 (Nov. 2012). ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-s16-s1. URL: http://dx.doi.org/10.1186/1471-2105-13-S16-S1.

[130] Jack Freestone et al. "How to train a post-processor for tandem mass spectrometry proteomics database search while maintaining control of the false discovery rate". In: (Oct. 2023). DOI: 10.1101/2023.10.26.564068. URL: http://dx.doi.org/10.1101/2023.10.26.564068.

[131] Wout Bittremieux et al. "A learned embedding for efficient joint analysis of millions of mass spectra". In: *Nature Methods* 19.6 (May 2022), pp. 675–678. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01496-1. URL: http://dx.doi.org/10.1038/s41592-022-01496-1.

[132] Samuel S Schoenholz et al. "Peptide-spectra matching from weak supervision". In: *bioarxiv* (2018). eprint: 1808.06576 (q-bio.QM).

[133] Varun Ananth et al. "A learned score function improves the power of mass spectrometry database search". In: (Jan. 2024). DOI: 10.1101/2024.01.26.577425. URL: http://dx.doi.org/10.1101/2024.01.26.577425.

[134] Tom Altenburg et al. "Foundation Model Enables Interpretable Open and Error-Tolerant Searching for Mass Spectrometry-Based Proteomics". In: (Dec. 2021). DOI: 10.1101/2021.12.01.470818. URL: http://dx.doi.org/10.1101/2021.12.01.470818.

[135] Dorte B. Bekker-Jensen et al. "An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes". In: *Cell Systems* 4.6 (June 2017), 587–599.e4. ISSN: 2405-4712. DOI: 10.1016/j.cels.2017.05.009. URL: http://dx.doi.org/10.1016/j.cels.2017.05.009.

[136] Michal Bassani-Sternberg et al. "Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry". In: *Nature Communications* 7.1 (Nov. 2016). ISSN: 2041-1723. DOI: 10.1038/ncomms13404. URL: http://dx.doi.org/10.1038/ncomms13404.

[137] Ann Cavers et al. "Behçet's disease risk-variant HLA-B51/ERAP1-Hap10 alters human CD8 T cell immunity". In: *Annals of the Rheumatic Diseases* 81.11 (Nov. 2022), pp. 1603–1611. ISSN: 0003-4967. DOI: 10.1136/ard-2022-222277. URL: http://dx.doi.org/10.1136/ard-2022-222277.

[138] Sarah M. Williams et al. "Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography–Mass Spectrometry for High-Throughput Single-Cell Proteomics". In: *Analytical Chemistry* 92.15 (July 2020), pp. 10588–10596. ISSN: 1520-6882. DOI: 10.1021/acs.analchem.0c01551. URL: http://dx.doi.org/10.1021/acs.analchem.0c01551.

[139] Dejan Burq Maximilien; Stepec. *Tesorai search*. V3. 2025. DOI: 10.17632/znkr2dm8fb.3.

[140] Wassim Gabriel et al. "Prosit-TMT: Deep Learning Boosts Identification of TMT-Labeled Peptides". In: *Analytical Chemistry* 94.20 (May 2022), pp. 7181–7190. ISSN: 1520-6882. DOI: 10.1021/acs.analchem.1c05435. URL: http://dx.doi.org/10.1021/acs.analchem.1c05435.

[141] Florian Meier et al. "Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer". In: *Molecular amp; Cellular Proteomics* 17.12 (Dec. 2018), pp. 2534–2545. ISSN: 1535-9476. DOI: 10.1074/mcp.tir118.000900. URL: http://dx.doi.org/10.1074/mcp.TIR118.000900.

[142] Bart Van Puyvelde et al. "A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics". In: *Scientific Data* 9.1 (Mar. 2022). ISSN: 2052-4463. DOI: 10.1038/s41597-022-01216-6. URL: http://dx.doi.org/10.1038/s41597-022-01216-6.

[143] Julian Lamanna et al. "Digital microfluidic isolation of single cells for -Omics". In: *Nature Communications* 11.1 (Nov. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-19394-5. URL: http://dx.doi.org/10.1038/s41467-020-19394-5.

[144] Jack Freestone, William Stafford Noble, and Uri Keich. "Reinvestigating the Correctness of Decoy-Based False Discovery Rate Control in Proteomics Tandem Mass Spectrometry". In: *Journal of Proteome Research* 23.6 (Apr. 2024), pp. 1907–1914. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.3c00902. URL: http://dx.doi.org/10.1021/acs.jproteome.3c00902.

[145] John Klimek et al. “The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools”. In: *Journal of Proteome Research* 7.1 (Aug. 2007), pp. 96–103. ISSN: 1535-3907. DOI: 10.1021/pr070244j. URL: http://dx.doi.org/10.1021/pr070244j.

[146] Bo Wen et al. “Assessment of false discovery rate control in tandem mass spectrometry analysis using entrapment”. In: (June 2024). DOI: 10.1101/2024.06.01.596967. URL: http://dx.doi.org/10.1101/2024.06.01.596967.

[147] Ana Marcu et al. “HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy”. In: *Journal for ImmunoTherapy of Cancer* 9.4 (Apr. 2021), e002071. ISSN: 2051-1426. DOI: 10.1136/jitc-2020-002071. URL: http://dx.doi.org/10.1136/jitc-2020-002071.

[148] Steven M. Yannone et al. “Toward Real-Time Proteomics: Blood to Biomarker Quantitation in under One Hour”. In: *Analytical Chemistry* 97.12 (Mar. 2025), pp. 6418–6426. ISSN: 1520-6882. DOI: 10.1021/acs.analchem.4c05172. URL: http://dx.doi.org/10.1021/acs.analchem.4c05172.

[149] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer Normalization”. In: *bioarxiv* (2016). eprint: 1607.06450 (stat.ML).

[150] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *bioarxiv* (2017). eprint: 1711.05101 (cs.LG).

[151] Felipe da Veiga Leprevost et al. “Philosopher: a versatile toolkit for shotgun proteomics data analysis”. In: *Nature Methods* 17.9 (July 2020), pp. 869–870. ISSN: 1548-7105. DOI: 10.1038/s41592-020-0912-y. URL: http://dx.doi.org/10.1038/s41592-020-0912-y.

[152] Mathias Wilhelm et al. “Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics”. In: *Nature Communications* 12.1 (June 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-23713-9. URL: http://dx.doi.org/10.1038/s41467-021-23713-9.

[153] Matthew C Chambers et al. “A cross-platform toolkit for mass spectrometry and proteomics”. In: *Nature Biotechnology* 30.10 (Oct. 2012), pp. 918–920. ISSN: 1546-1696. DOI: 10.1038/nbt.2377. URL: http://dx.doi.org/10.1038/nbt.2377.

[154] Piero Giansanti et al. “An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas”. In: *Cell Reports* 11.11 (June 2015), pp. 1834–1843. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.05.029. URL: http://dx.doi.org/10.1016/j.celrep.2015.05.029.

[155] Hendrik Weisser and Jyoti S. Choudhary. “Targeted Feature Detection for Data-Dependent Shotgun Proteomics”. In: *Journal of Proteome Research* 16.8 (July 2017), pp. 2964–2974. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.7b00248. URL: http://dx.doi.org/10.1021/acs.jproteome.7b00248.

[156] Fengchao Yu, Sarah E. Haynes, and Alexey I. Nesvizhskii. “IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs”. In: *Molecular amp; Cellular Proteomics* 20 (2021), p. 100077. ISSN: 1535-9476. DOI: 10.1016/j.mcpro.2021.100077. URL: http://dx.doi.org/10.1016/j.mcpro.2021.100077.

Acknowledgements

First of all, I would like to thank my family, who taught me to be ambitious and hardworking. Without you, this would have been just a dream.

To Leonor, whose kind heart shines through in everything she does—thank you for your unwavering support.

To my friends from Colombia: we grew up together and have shared so many unforgettable moments that I can't imagine my life without you. You've made this achievement possible, and I feel incredibly fortunate to have you in my life.

A special acknowledgment goes to my colleagues in the Cox Lab. Thank you for always being willing to help, even when it wasn't easy. To Dr. Juergen Cox, thank you for giving me the freedom and guidance necessary to develop this project, and for supporting me beyond academia.

To Prof. Petra Schwille, thank your generous support, which made my doctorate possible.

Thank you to my friend Dr. Christoph Wichmann—you made a significant difference in my Ph.D. through your insightful discussions, valuable ideas, and all the meaningful conversations.

To all the wonderful people I've met in Germany, a big thank you for being part of this journey.