Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Application of Machine Learning and Deep Learning for Proteomics Data Analysis

Shivani Tiwary

aus
Bhabhua Kaimur, India
2021

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Jürgen Cox betreut und von Herrn Prof. Dr. Thomas Carell von der der Fakultät für Chemie und Pharmazie vertreten.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 13.01.2021

Shivani Tiwary

Dissertation eingereicht am:	13.01.2021
1. Gutachter:	Prof. Dr. Thomas Carell
2. Gutachter:	Dr. Jürgen Cox
Mündliche Prüfung am:	19.02.2021

Summary

A diverse set of both supervised and unsupervised methods such as hidden Markov models, neural networks, support vector machines, Bayesian analysis, and clustering algorithms have been applied over the last years to biological data analysis. There is a strong dependency between the amino acid sequence of a protein and its biological properties, and determining these properties such as localization, structure and function given a biological sequence, is one of the greatest challenges in computational biology. Since 2006, deep neural architectures have become popular. Deep learning was successfully used in many domains such as speech and image recognition and natural language processing tasks. Methods such as SVMs, random forests and neural networks with a single hidden layer require careful design of features so that patterns can be learned by the algorithms. In contrast, deep neural networks have been shown to outperform these conventional methods in some areas as they are capable of learning intermediate representations, where each layer is an abstract representation based on the abstractly represented features of the previous layer. Deep learning algorithms have already shown some promising results in genomics and proteomics fields[1,2].

Mass spectrometry-based proteomics experiments provide data that helps in accurately identifying and quantifying proteins in biological samples of interest. In bottom-up proteomics, peptide identification by fragmentation resulting in MS/MS spectra is the fundamental approach. The fragmentation chemistry is still not understood completely and could theoretically be solved using quantum chemistry. Alternatively, machine learning based prediction can be applied. It has been used in the past to predict spectrum intensity with limitations such as models not being independent of variable length of peptide sequences, and separate model creation for different fragment ions, charges and fragmentation type. Feature space were designed using biophysical chemical properties of amino acids, and properties of mass spectrometry instruments.

The aim of this thesis, is to develop a regression model, which predicts fragment spectrum intensities taking peptide sequence as input and to provide proof of concepts of benefits of the spectrum intensities in both data dependent and data independent acquisition data analysis. ***Article 1***, in collaboration with Verily life sciences we developed two regression models to predict intensities. *DeepMass:Prism*, a bi-directional long short

term memory (LSTM) model trained on 60 million peptide spectra from publicly accessible datasets, which captures sequence features that contribute to peptide fragment-ion abundance. *wiNNer*, a fully connected neural network model based on a sliding window approach, where the feature space is centered around the target bond for which prediction is done. Both the models overcome the limitation of covering peptides of variable lengths.

Results show that *DeepMass:Prism* can successfully predict MS/MS spectrum intensities nearly as accurate as technical reproducible intensities. *wiNNer* has slightly inferior predictive performance but it is easily re-trainable on smaller training dataset and is computationally inexpensive. The predicted spectrum as shown in ***article 1*** can benefit analysis of both, data-dependent acquisition and data-independent acquisition. In data-dependent acquisition (DDA) approach spectra are identified using database search engines by matching the experimental MS/MS spectra with the theoretical spectra generated from the protein databases(e.g.Uniprot[3]). MS/MS spectra intensity information could be of high relevance in correctly identifying the peptide sequence. However, it is not used by any search engines. In the ***article 1***, predicted intensity was integrated into the peptide score calculation in the Andromeda search engine and we demonstrated an increase in the total number of peptide identifications as a function of q-value.data-independent acquisition (DIA), which depends on sample specific spectral libraries generated by DDA experiments to identify peptide, which makes it cost and time effective. In the ***article 1***, spectral libraries generated from DDA experiments were replaced by in-silico spectral libraries using *DeepMass:Prism* showing highly correlated peptide abundance quantification.

The study in ***article 2*** provides important insights into the evolutionary relationships between H. antecessor and other hominin groups. The authors used enamel proteomes to investigate hominin biology across the existence of the genus Homo. To validate the enamel peptide spectrum matches, the *wiNNer* algorithm was used to predict MS/MS spectrum intensity. For predictions, *wiNNer* was trained on randomly cleaved and heavily modified peptides from the ancient samples. The results show that the *wiNNer* model trained on heavily modified peptides provides a predictive performance similar to that of the *wiNNer* model trained on modern, trypsin-digested samples, assuring

accurate sequence identification for the phylogenetically informative peptides (median Pearson correlation coefficients of 0.76).

The PRoteomics IDEntifications (PRIDE)[4] database is one of the world's largest mass spectrometry-based proteomics data repositories to deposit proteomics experimental data. PRIDE supports data deposition, automatic and manual curation of related experimental metadata, to promote and facilitate the reuse of public proteomics datasets. It also has the quality control pipelines and visualization components to enable the assessment of the data quality. To support handling of the data Proteomics Standards Initiative (PSI) created specific data standard formats such as mzTab, mzIdentML. In ***article 3***, the authors discuss recent developments and improvements in the PRIDE resources and the tools they used. The thesis also covers the mzTab table generated in MaxQuant[5] for the complete submission in PRIDE repository.

The protein sequence features such as disorder regions[6] and low complexity regions[7] makes the protein structure unstable and causes aggregation of proteins. The proteins form nuclear aggregates and can cause various neurodegenerative disorders such as amyotrophic lateral sclerosis and Huntington's disease. The authors in ***article 4*** used a combination of methods such as fluorescence imaging and proteomics to investigate the aberrant proteins in the nucleus focusing specifically on the role of the nucleolus and its phase-separated nature in protein quality control. The results showed that the nuclear proteins where highly enriched in disordered as well as low complexity regions causing in misfolding of the proteins.

Contents

List of Figures

Chapter 1

Introduction

Rapid technology development of mass spectrometer instruments in conjunction with advanced bioinformatics analysis capacities now allows in-depth analysis of proteomics samples. Proteins are a functional entity in cells and are involved in the structure, function, and regulation of cells, tissues, and organs. The common workflows used in proteomics are the shotgun (bottom-up) approach (shotgun)[8] and top-down approach[9–12]. The shotgun proteomics workflow begins with taking the sample of interest(e.g.protein extracted from cells or tissues) and digesting the protein to get peptides. This is often followed by peptide fractionation and enrichment, before the separation of peptides by high performance liquid chromatography (HPLC). Ionized peptides are then passed through a high-resolution mass spectrometer, peptide isotope patterns are recorded from the full (MS^1) spectra, peptide precursors are selected for fragmentation, and fragment (MS^2) spectra are recorded. Lastly, software like MaxQuant[13], Mascot[14], Sequest[15], and X! Tandem[16] are used to identify and quantify peptides, proteins, and post-translational modifications[17].

Later downstream statistical data analysis is performed, for example gene ontology enrichment or network analysis to the results to get the gene ontology enrichment or network analysis, to understand proteins and their function. Perseus, a user friendly software can be used for the downstream statistical data analysis[18]. With the large amounts of data generated by high throughput instruments, it is possible to use machine-learning algorithms to reveal features and patterns from protein and peptide sequences, to solve biological problems like protein folding and to understand the function of the proteins. Deep learning algorithms, with successful applications in speech

recognition and image analysis, harbor great potential to understand and predict mass spectrometry data. The following introduction section is divided into four subsections a) mass spectrometry-based proteomics, b) advances in machine learning algorithms, c) protein sequence features, and d) application of machine learning in proteomics specific to spectrum predictions.

1.1 Mass spectrometry-based proteomics

The complete set of proteins that are produced or modified by the organism is known as the proteome. Proteomes are the protein complements of genomes, and they are highly dynamic and interact with other proteins and biomolecules. After the successful completion of the human genome project, we now know that there are more than 20,000 genes in human, which lies between chicken and grapes[19]. The number of genes does not indicate the complexity of organisms. Moreover, the cells and tissues of a single organisms have the same set of genes yet completely different physiology and functionality. Regulation of these genes, their translation into proteins, and the modifications, localization, and complex structure of the proteins, generate these physiological and functional differences. post-translational modifications (PTMs)[20] and splice variants[21] increase the complexity of the proteome within individual cells[22,23].

The large scale study of the proteome is defined as proteomics, a term coined by Marc Willing in 1994. Proteomes are identified and quantified using mass spectrometry-based technology. The mass spectrometer was developed to determine the mass of proteins, but it needs an efficient ionization method, which was not available until decades later. In the late 80s, two ionization methods electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) were developed, capable of analyzing proteins and later earning the Nobel prize[24,25]. One of the advantages of ESI compared to MALDI is that proteins or peptides are ionized in the liquid phase. Hence, ESI can be directly coupled to liquid chromatography and allows the analysis of complex protein and peptide mixtures.

With the advancement of the MS-based method, we aim to study thousands of proteins in a cell or tissue and their post-translational modifications using complex experimen-

tal designs. These studies aim for the identification of more than 10,000 proteins in a system[26]. Some other applications of proteomics are to understand biological function such as protein-protein interactions, substoichiometric protein modifications, and cellular localization using isolation and enrichment strategies applied during sample preparation. UniProt (universal protein resource)[3] is a comprehensive, high quality, and freely accessible resource of protein sequences and functional information, which is helpful for the identification of proteins using MS-based methods. Peptide-based shotgun proteomics (bottom-up approach) is the most commonly used method in protein identification and quantification, and the workflow is discussed in detail in the following sub-sections (see Figure 1.1).

1.1.1 Sample preparation

In shotgun proteomics, protein is first extracted from the sample of interest (e.g. cell, tissue, or plasma) using proteolytic digestion followed by cleaning the sample using detergent, and followed by the enzymatic digestion of protein into peptides typically using trypsin (see Figure 1.1). The major steps in the sample preparation are as follows. Proteins are extracted from the biological material by mechanical disruption and/or detergent based lysis. Depending on the experimental study, either all proteins are denatured, or native states are preserved using physiological buffers with mild detergents. For complete denaturation of proteins and to dissolve lipids, detergents like 4 % SDS or triton are used. Sonication, bead-milling, rotor start, blending, and heating of the sample can also be performed together with lysis. Before enzymatic digestion, detergents are removed from the samples. The detergents are not MS-compatible because they co-elute with peptides and usually ionize well with electrospray, and thus can cause misidentification. Acetone or ethanol precipitation of proteins or membrane-based cleanup (FASP)[28] can be employed if the protein amount is low. For a higher amount of proteins, it is usually advisable to use MS-compatible detergents or chaotropic agents for lysis.

After the proteins are cleaned, they are digested into peptides using proteases such as trypsin, LysC, and chymotrypsin. Trypsin cleaves C-terminal to arginine and lysine (if not followed by proline)[29]. In acidic conditions, the resulting peptides have at least two positive charges, one at the N-terminus and the other at the side chain of the terminal

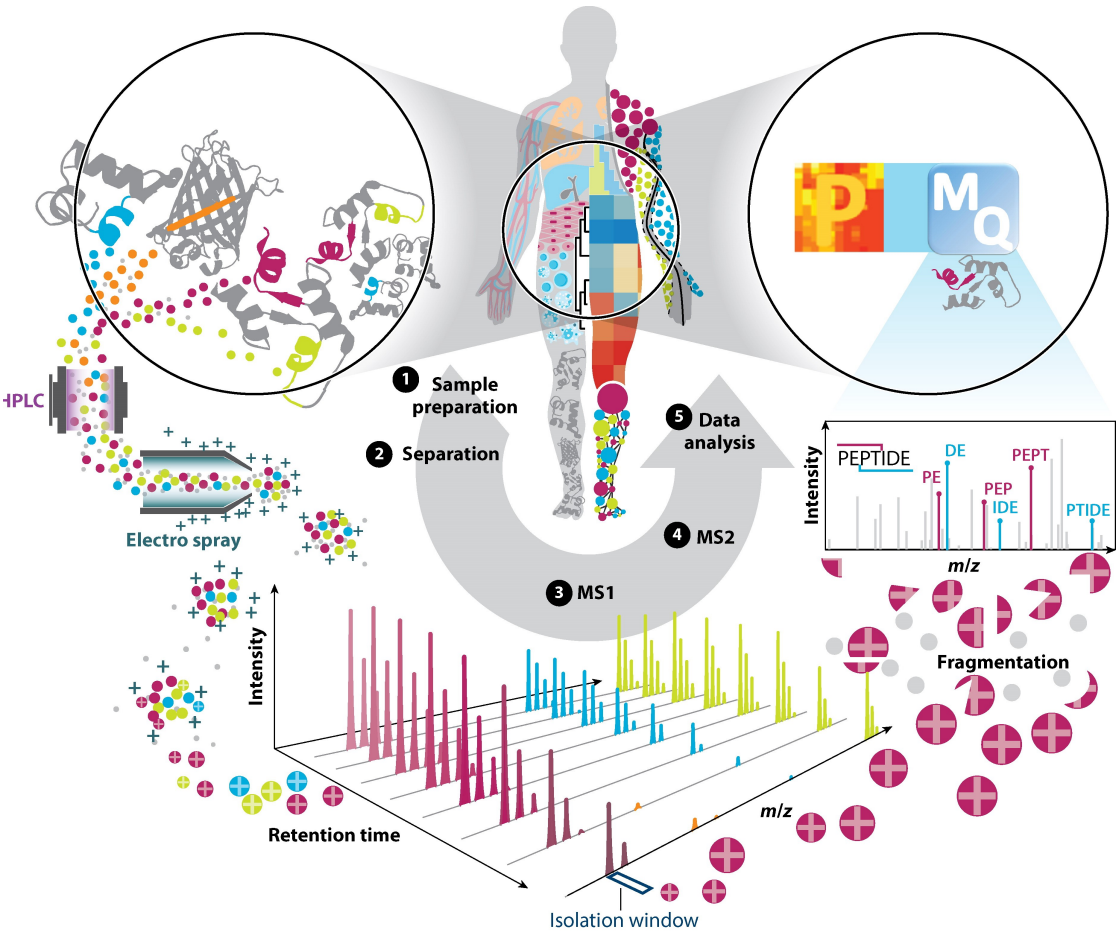


Figure 1.1: Shotgun proteomics workflow begins with sample preparation, HPLC separation and ionization of peptides, ionized peptides injected into high resolution mass spectrometer, MS^1 spectra containing peptide isotope patterns are recorded, peptide precursors are selected for fragmentation and fragments MS^2 spectra are recorded. MS^1 and MS^2 spectra are then analyzed by computational proteomics software. Taken from the review [27].

residue. The additional charges enable peptides to be distinguished from non-peptide contaminants. The distribution of lysine and arginine in proteins generates on average 10 amino acid long peptides, which makes the length suitable for the high-resolution analysis in the commonly used mass analyzer. The lysine specific enzyme LysC is also widely used. It is active at 8M urea and has a higher efficiency than trypsin in cleaving C-terminal to lysine[30]. Chymotrypsin cleaves C-terminal to aromatic residues and GluC-D[22] cleaves C-terminal to aspartate and glutamate. They are mostly used to increase peptide coverage of proteins or to generate peptides with different properties. Earlier, proteins were digested 'in-gel' after separating them on an SDS polyacrylamide gel. Currently, 'in-solution' digestion is the method of choice, especially in combination with HPLC.[31,32]

1.1.2 Chromatography

In-solution digestion of the complex proteome might yield 100,000 unique peptides[33] and these peptides need to be separated efficiently. In chromatography systems, the analytes (peptides) differentially interact with the stationary phase of strong cation exchange (SCX) or reversed-phase (RP) chromatography columns due to their different physiochemical properties[8]. RP chromatography is based on the hydrophobic interaction between peptides and the C18-silica of the column. By applying a pH gradient to the mobile phase, all peptides are eluted from the column throughout a mass spectrometry (MS) run. The resolving power of a chromatography[34] method can be optimized by changing the column length[35]. A longer column allows for more interaction between peptides and the stationary phase, which in turn increases the resolution. The smaller inner diameter of the column and uniform particles as filling material helps in increasing chromatography resolution as it reduces the number of flow paths (eddy diffusion) and the negative effect of mass transfer. Increasing the gradient length increases the resolution. Longer gradients can cause peak broadening and consequent reduction of the ion current due to dilution. But it is improved with higher flow rates, which causes higher backpressure and reduces ionization efficiency[36]. With the ultra-high pressure(up to 100,000 bar) and column heating device, the chromatographic performance has significantly improved.

1.1.3 Mass spectrometer

The mass spectrometer has three central parts, the ion source, the mass analyzer and the detector. The ion source and mass analyzer will be discussed briefly in the next sections.

Ion Source

The ion source ionizes the particles, and these ionized particles then enter the vacuum of the mass spectrometer. Until the 1980s, the study of proteins or peptides was incompatible with MS as they could not be transferred into the vacuum of the mass spectrometer without being destroyed. Introduction of two ionization methods, ESI and MALDI solved the problem. The methods shared the 2002 Nobel prize in Chemistry. MALDI[24] creates ions by pulsing the sample loaded onto a solid matrix with a laser. The laser pulse excites the matrix molecules, which leads to its desorption along with ionized analyte molecules, and the mass is then measured in a time-of-flight(TOF) analyzer. Differently to MALDI, electrospray disperses a stream of liquid into a charged aerosol when high voltage is applied to the emitter[25,37,38]. The soft-ionization technique enables the analysis of intact proteins and peptides from solution, which makes it attractive for liquid chromatography (LC)-MS analysis. ESI yields multiple charged peptide ions with one charge per kDa. Therefore, even large masses are recorded in narrower $^{m}/_{z}$ range[38].

Mass analyzer

There are different types of mass analyzers used in proteomics and they are described briefly here. The mass analyzers can be broadly classified into two types, trap-based analyzers and beam-based analyzers. Trap-based analyzers include 3D and 2D ion traps (linear ion trap quadrupole, LTQ), Fourier transform ion cyclotrons resonance (FT-ICR)[39], and the Orbitrap analyzer[40]. The beam-based analyzers are made up of 2D quadrupole[41] and the time of flight (TOF) instruments that continuously scan incoming ions. The performance of these analyzers can be described by parameters like mass resolution, mass accuracy, scan speed and sensitivity.

Mass resolution: High mass resolution should be able to distinguish two peaks with sim-

ilar m/z. Ion traps and quadrupoles have a low resolution (∼1000), TOF instruments perform better (>10,000). However, the highest resolving power is provided by FT-ICR and Orbitrap analyzers (>100,000). Orbitrap is the most preferred high-resolution analyzer.

The *mass accuracy* is the deviation between the theoretical mass and the experimentally determined mass and is read as the mass error in parts per million or billion (ppm, ppb). The mass accuracy depends on many parameters such as resolution and signal to noise ratio. The mass error can be corrected using internal and external calibrations.

The *dynamic range* shows how well the mass analyzer detects low abundant ions together with very high abundant ions.

The *scan speed* indicates how fast the m/z range can be monitored and for many instruments, it is inversely correlated with the resolution. FT-ICR analyzers are usually slowest, ion traps and Orbitraps are comparably faster, and quadrupoles and TOFs are the fastest.

Sensitivity parameters show the detection limit of a mass analyzer. Standard ion traps and linear ion traps have electron multipliers as detectors, which are capable of detecting single ions and are thus highly sensitive. FT-ICR detectors usually require more charges to distinguish a signal from noise. In the Orbitrap analyzer, single ion detection is possible due to improved electronics and thermal stability[42].

The **Orbitrap** analyzer was introduced in 2000 by Alexander Makarov[43]. It uses an electrostatic field, which is used in quadrupoles, or 3D or 2D ion traps, around a cone shaped electrode. The electrostatic field generates a quadro-logarithmic potential distribution and is composed of a quadrupole field, generated by the ion trap outer barrel-like electrodes and the field of the spindle-like inner electrode. Over the years, Orbitrap performance has improved significantly because of the development of enhanced Fourier transformation and by implementing a smaller, high-field Orbitrap analyzer. The mass-to-charge ratio m/z of the injected ions is measured in the mass analyzer. The transients of ions oscillating inside the trapping analyzer are recorded and transformed into m/z values using Fourier analysis. The Orbitrap has many favorable characteris-

tics for lower-mass peptide analytes, such as very high resolution and mass accuracy. Many modern mass spectrometers, such as the Q-Exactive HF operate multiple mass analyzers in tandem. Quadrupoles are used for the selection of ions within a specified m/z range and traps are often utilized for the accumulation of ions prior to mass analysis. Finally, the ions reach the detector, which counts the number of ions observed at each m/z value.

1.1.4 Ion Fragmentation

Tandem mass spectra[16] generates many fragment ions and a detailed interpretation of the most abundant peak is required for confident peptide assignment. Peptides are fragmented in the MS and each amino acid (aa) residue has different ion fragmentation capabilities. Mass spectrometric fragmentation, known as MS/MS or MS^2, is used to get information about the sequence of each peptide. Selected peptides of interest are isolated within a desired m/z window, subjected to fragmentation and the fragments are measured in a subsequent mass spectrum[44,45]. The peptide fragmentation is done by

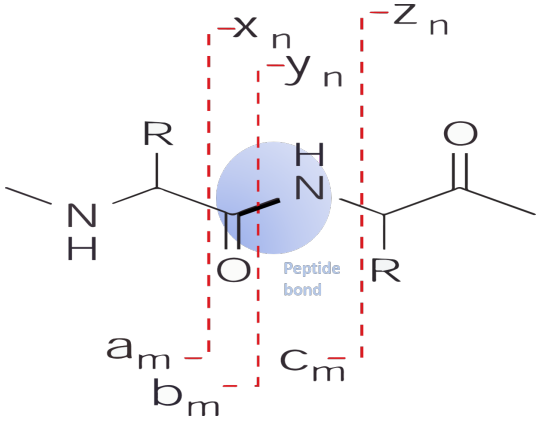


Figure 1.2: The backbone bonds cleave into six different types of fragment ions. The N-terminal fragment ions are a, b or c, while the C-terminal fragment ions x, y or z. The subscript n and m is the number of amino acid residues. Adapted from[46]

inducing dissociation of the peptides by collision with an inert gas such as He or N_2. The kinetic energy is partially converted into internal energy, which breaks the chemical bonds. The collision energy required for efficient fragmentation depends on the peptide mass and charge state. This generates sequence-specific backbone fragments referred to as ions. The most common fragmentation methods are collision-induced dissociation

(CID) and Higher-energy collision dissociation (HCD)[47], which fragment peptides at the amide bond to a series of b- and y- ions, N- or C- terminus, respectively (see Figure 1.3). A full series of either b- or y-types ions in principle, allows the entire amino acid

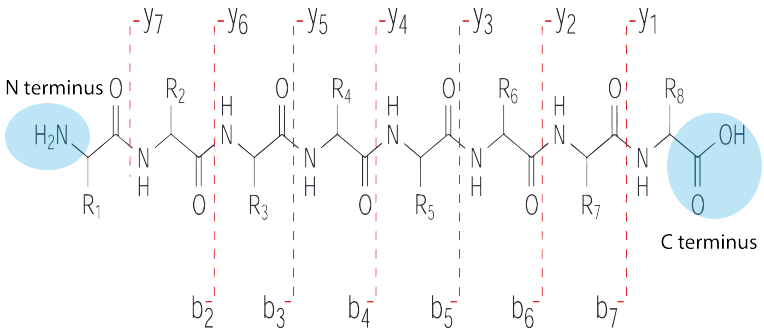


Figure 1.3: Example of complete y-and b-ion series. The y-ions are numbered consecutively from the C-terminus, the b-ions are numbered consecutively from the N-terminus. The difference between consecutive ions gives the masses of the corresponding amino acids. Adapted from[46]

sequence to be read from a fragment ion spectrum[48]. Neutral losses of molecules such as NH_3 and H_2O from fragments ions can complicate tandem mass spectra. The b-ions are chemically less stable and often further fragmented, leading to a prominent y-ion series in HCD. While ion trap CID fragmentation spectra are usually recorded at low resolution and low mass accuracy[49,50], HCD usually features high resolution and high mass accuracy[51]. The backbone fragmentation are a,b,c for N-terminal and x,y,z for C-terminal types, depending on the cleavage position on the peptide backbone[52,53] (see Figure 1.2). The activation of the peptide with an electron such as in Electron-capture dissociation (ECD) and Electron-transfer dissociation (ETD), breaks the N-C bond and generates c- and z- ions[54,55]. The advantage of ETD and ECD is to analyze intact proteins and peptides carrying PTMs where one needs to avoid fragmentation of weak bonds. The orthogonality of ETD/ECD compared to HCD/CID can be very useful to increase the fragmentation. Peptide fragmentation is not clearly understood and the mobile proton model is the most accepted framework to understand the dissociation process.

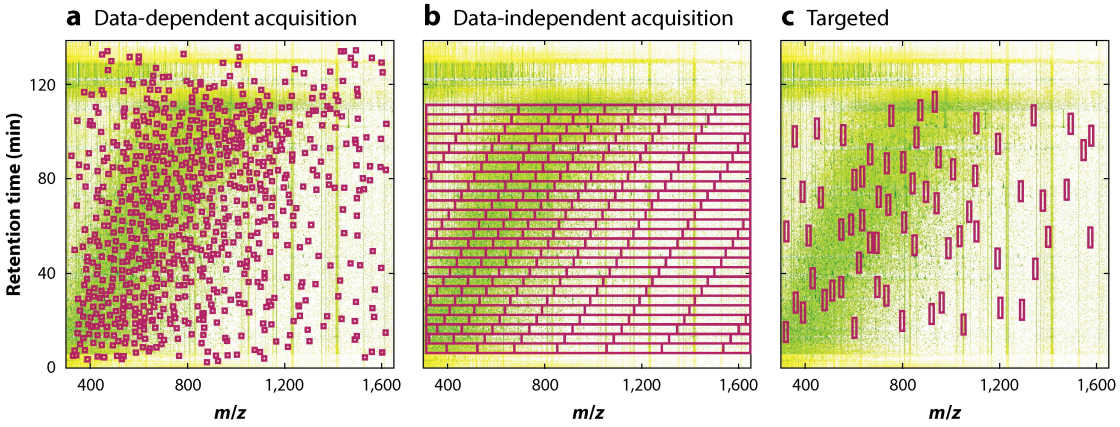


Figure 1.4: Data acquisition modes in bottom-up proteomics. (a) The most common acquisition mode in bottom up proteomics is DDA, where at given retention time, top n most intense peptide features are selected for fragmentation (b) In DIA set of constant mass ranges are isolated for fragmentation. (c) A list of peptides is targeted for fragmentation based on the peptides of interest. Taken from[27].

1.1.5 Acquisition methods

There are three different types of acquisition mode 1. targeted acquisition, 2. Data dependent acquisition and 3. Data independent acquisition (see Figure 1.4). The most popular acquisition technology in shotgun proteomics is data dependent acquisition. In targeted mode, the mass spectrometer is configured to target a predefined set of masses, aiming for the highest possible quantitative accuracy and reproducibility. DDA and DIA are discussed in the following sections.

Data dependent acquisition

Mass spectrometers can be operated in a number of different acquisition modes, which determine the succession of MS^1 and MS^2 scans during a measurement run (see Figure 1.4). DDA relies on the observed peaks on the MS^1-level to decide which ions will be subsequently isolated, fragmented and sent for MS^2 analysis. The goal of the MS^2 analysis is to sequence peptides by measuring their fragment ion series. To this end, fragmentation energies are optimized to induce a single peptide backbone breakage that gives rise to a set of complementary fragment ions. Time constraints do not allow for the exhaustive sequencing of peptides. Instead, a common strategy is that eluting

peptides are measured in a survey scan (MS^1 or full scan) followed by selection and fragmentation of the top N most abundant peptides that were not fragmented before. The survey scan usually covers a m/z range between 300 to 1650 Th at a resolution of 60,000. After each MS^1, five to twenty of the most intense features with a charge state higher than one are sequentially subjected to fragmentation, and the fragment masses are recorded in separate MS^2 spectra. The acquisition cycle in the Q Exactive is 1 second, which consists of MS^1 and 15 MS^2 scans. To prevent re-sequencing of the same peptide, precursors with the same mass are excluded from selection for 20 to 40 seconds. If the sample is very complex, under sampling can occur, which gives rise to missing value problem in DDA, where some peptides are sequenced in one but not in identical samples. DDA performance is compromised when the sample becomes complicated because of the semi stochastic selection of precursor ions which limits both identification and quantification.

Data independent acquisition

With advances in instrumentation and software[56], SWATH DIA has emerged as an alternative to DDA for proteomics analysis. After acquiring the MS^1 scan, the entire mass range is segmented into overlapping windows. Subsequently, each mass window is fragmented and a MS^2 scan is obtained, regardless of the measured MS^1 information. DIA-UMPIRE[57]: In this acquisition method the mass spectrometer selects a precursor range of about 10 to 25 m/z units and cycles through the mass range. Methods like SWATH MS have advantages like all the precursors are sequenced but with reduced dynamic range. The resulting MS^2 are very complex for each isolation window and elution time, so a peptide fragmentation library is required to identify the peptides. DIA data interpretation is more complex but the SWATH MS quadrupole-TOF instruments[58] are fast enough to sample the mass range as the time window is smaller than the average time for peptide elution. But the library generated for peptide identification is from DDA methods and is time taking and cost effective. It overcomes the limitation of missing values in DDA. DIA uses co-elution and co-fragmentation. DIA avoids the detection and selection of individual precursor ions during LC-MS analysis and just fragment everything in a window and it generates very complex spectra but you do not lose any ions at any time[59].

1.2 Computational mass spectrometry

1.2.1 Peptide identification

Peptide sequences are identified from MS^1 and MS^2 fragmentation spectra using search engines like Andromeda and Mascot. Most popular search engines use a database search approach[14–16,60,61], using protein databases from Uniprot[3] or Ensemble[62]. The protein sequences in the database are digested *in silico* into peptides following the cleavage rule of the proteases used in the experimental design (e.g. trypsin). For each *in silico* peptide, a list of expected fragment masses is generated based on the fragmentation method used in the experiment (e.g. HCD, CID). For each experimental spectrum, the search engine calculates the match score against all the theoretical MS^2 spectra within a specified peptide mass tolerance. The highest scoring peptide spectrum match (PSM) is a candidate to identify the peptide. However, these highest scoring PSMs can still be false positives, so it is necessary to control the false discovery rate (FDR) using a target-decoy approach[63]. In this approach, experimental spectra are searched against the target database and also against the decoy database. The decoy database contains reversed amino acid sequences of target sequences, which do not occur in nature. In reversed sequence, the lysine and arginine (e.g. trypsin digestion) are swapped with the preceding amino acid to avoid the exact same mass for forward and reverse peptide while preserving the context of each amino acid[64]. Spectra are then matched to this combined target-decoy database, which is designed to produce false-positive PSMs. Comparing score distribution of target and decoy PSMs, posterior error probabilities are calculated and FDR is controlled[5]. Additional peptide features besides the search engine score, such as length of the peptide and number of missed cleavages help in distinguishing the true identification from false positives. Tools such as PeptideProphet[65,66] and Percolator[67] use linear discriminant analysis or support vector machines (SVM) to get the correctly identified peptide. To further improve identification and support database scoring, machine learning was used to predict spectra intensity[68,69], but failed to improve upon the state of the art. De novo peptide identification[2] using deep learning yielded improvements in identifications. De novo peptide sequencing is another approach to identify peptides from fragment spectra. There are many existing tools that identify peptides using only information from input spectra and the characteristics of the fragmentation method, some also use homology sequencing[70,71].

All identified peptides are then assembled into proteins. Proteins upon digestion can have many peptides, whereas one peptide can originate from one or more proteins. Proteins that are identified by unique peptides are assembled into individual proteins. Proteins that are not discriminated by unique peptides are combined in protein groups. Longer peptides are more likely to be unique and more informative. Peptides of length 7 or longer are expected to be informative and useful. The Parsimonious model[72,73] applies Occam's razor principle to the protein inference problem by finding a set of proteins that is as small as possible to explain the observed peptides. Statistical models[74] can assemble large amounts of weak peptide identifications[27]. Each protein group contains a set of proteins that cannot be distinguished from each other based on the observed peptides. Either the proteins in a protein group have equal sets of identified peptides or the peptide set of one protein is a proper subset of that of another protein. Assembled proteins are also FDR controlled based on ranked protein posterior error probability (PEP), which we get from the product of respective peptides PEPs[5,61]. It is important to limit the false positive proteins present in the sample, as it impacts the biological outcome of the relevant study.

1.2.2 Quantification methods

In addition to protein identification, protein quantification makes proteomics the most powerful tools in biological processes. In proteomics there are two level of quantitative information. Relative quantification that measures the difference between same protein in two or more samples. The absolute quantification determines the absolute amount of proteins within a sample, by determining copy numbers or concentration per cell[75](see Figure 1.5). In relative quantification, quantitative ratio of protein concentration between the samples are calculated.stable isotope labeling by amino acids in cell culture (SILAC), tandem mass tags (TMT) and isobaric tags for relative and absolute quantification (iTRAQ) are some popular quantitative methods. The isotopic labels are done in two ways: metabolic labeling and chemical labeling.In metabolic labeling, the stable isotopes are introduced in living cell or organism through its metabolism. Example of metabolic labeling are SILAC, CTAP, Neucode. In chemical labeling techniques, the stable isotopes are added in chemical reaction during sample preparation. SILAC[76,77] is one of the most popular methods for quantitative proteomics that detects differences in protein abundances among the samples. It uses non-radioactive isotope

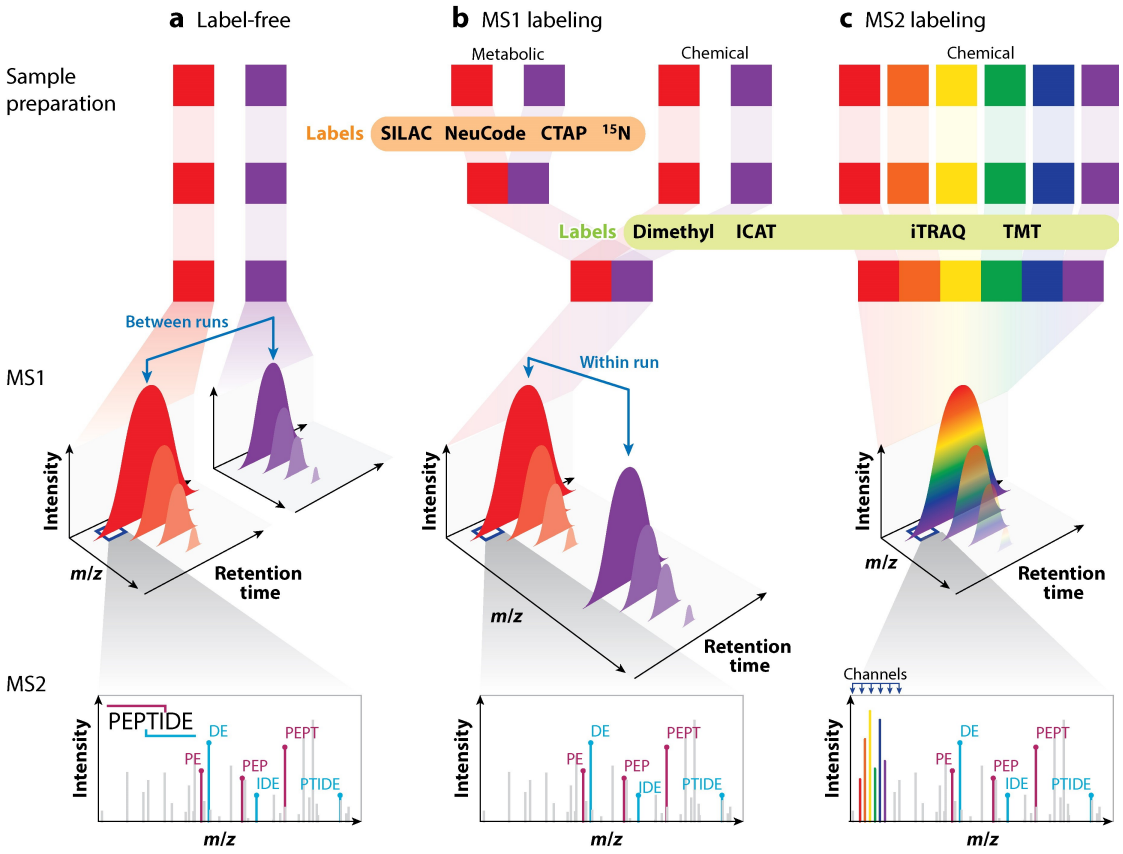


Figure 1.5: Label-free, metabolic, or chemical labeling approach for relative quantification. (a) In label-free, the quantification is done for each peptide feature between extracted ion chromatograms in different LC-MS runs. (b) SILAC, dimethyl, NeuCode are MS1 based label quantification, where multiple samples will appear as differentially labeled isotope patterns in the MS1 spectra. (c) iTRAQ, TMT are MS^2 based label quantification, where signal appear as reporter ions in the low-mass range of the MS^2 spectra. Taken from[27].

labeling with heavy isotopes mostly ^{13}C, ^{15}N, which gives light control and a heavy sample. Introduction of these isotopes does not change physico chemical properties of peptides, but only their mass. Hence, they behave exactly as the natural counterparts in the cell, during sample preparation and during HPLC separation. With difference in mass, it is easily distinguished in MS measurement. After labeling step, samples can be mixed and analyzed together in LC-MS/MS run. In isobaric labeling different samples are labeled with different molecules per sample. The molecules have same mass but eject different reporter ions upon fragmentation. The advantage of isobaric labeling is its multiplexing capacity. Tandem mass tag[78] can now multiplex up to 16 channels.

Label free quantification

Protein quantification without isotopic labels has several advantages. For example, it is applicable to any sample and to materials that cannot be metabolically labeled like clinical samples. There is no limit on number of samples for comparisons. Label free method are the simplest approach. However, it requires a controlled workflow as robustness and accuracy of quantitative information is reduced. Earlier label-free quantification methods were based on correlation of mass spectrometric signal of peptides with the relative or absolute quantity[79,80]. Spectral counting the abundance of proteins is estimated by the number of MS^2 spectra recorded for each protein[81]. High mass resolution, accuracy and high peptide identification rates are crucial for accurate quantification in both isotope-label-based and label-free methods. More recently developed label-free strategies make use of high resolution data and employ the MS^1 ion intensities of all the identified peptides (extracted ion current or XIC) to extract quantitative protein information. The MS^1 intensity is directly proportional with the number of ions, within the linear dynamic range of the instrument. The MaxLFQ[82] algorithms calculate ratios of normalized peptide intensity. MaxLFQ uses MS^1 intensity, sometimes also includes intensities got from matching between the runs and outputs relative abundance profiles over multiple sample.

1.3 Advances in machine learning algorithms

Machine learning has been successfully used in different research areas like structural biology and proteomics. Proteomics projects usually have as input the amino acid se-

quence of peptides or proteins and the algorithm learns the properties of the sequence given the numerical feature matrix. First-generation neural networks date back to the 1960s, with the introduction of perceptron by Frank Rosenblatt in 1962. Selective features were provided with weights and objects were recognized by learning through those weights. However, one of perceptron's biggest limitations is that it can only learn linearly separable patterns. Second generation neural networks were using Back propagation (BP), which became popular around 1985. BP error is used in combination with an optimization method such as gradient descent. The training algorithm involves an iterative procedure for minimization of an error function, with adjustment to the weights being made in a sequence of steps. Unfortunately, BP based training of deep neural networks, failed to optimize these weights and reduce error with many hidden layers. BP gets very slow in networks with multiple hidden layers. Sometimes BP can also get stuck in poor local optima when the batch mode or even stochastic gradient descent BP algorithm is used. They were not optimal for deep networks. Later, for a decade, there was a slight diversion from deep learning to shallow learning algorithms. In the 1990s, Vapnik and his coworkers developed a very clever type of perceptron called SVM[83], but with the same limitations as perceptron, and it was used only for labeled and linearly separable data. With the success of SVM and other machine learning methods, many researchers abandoned neural networks with multiple adaptive hidden layers. Later shallow-structured architectures such as Gaussian mixture models (GMMs), conditional random fields (CRFs), maximum entropy (MaxEnt) models, SVM, logistic regression, kernel regression, and multilayer perceptrons (MLPs) with a single hidden layer including extreme learning machines (ELMs), gained extreme popularity in different research areas. With the advent of graphical processing units (GPUs), the mathematical calculations became very fast. Self-driving cars using deep learning algorithms are one of the future applications and it is already being used by Tesla. AlphaGo[84] a deep learning model, which defeated human, was one of the first success stories and marked the beginning of an era where deep learning is being used in all domains.

Deep learning is a field in the machine learning research community introduced by Hinton et al, Bengio and Le Cunn around early 2000s. It models high-level abstraction in the data by using model architectures composed of multiple non-linear transformations. Deep learning is motivated from the deep architecture of the human brain.

The human brain organizes ideas and concepts hierarchically by first learning simpler concepts and then composing them to represent complex ones. Likewise, deep learning has multiple levels of abstraction and processing. Deep learning had immense success in a number of traditional artificial neural network applications, such as image recognition[85] and speech recognition. In this chapter, classical machine learning is discussed and its applications in biology. Later, detailed descriptions of recurrent neural networks its adaption, and different types of models, are discussed and some of its applications and how it has potential to be used for biological sequence data to predict mass spectrometry data. Deep learning automatically extracts features in each hidden layer thus does not require manual feature extractions. To use a simple example, a deep neural network tasked with interpreting shapes would learn to recognize simple edges in the first layer and then add recognition of the more complex shapes composed of those edges in subsequent layers. DeepBind[86] and DeepSEA[1] are examples of deep learning used in biological datasets. DeepNovo[2] is used for de novo peptide sequencing.

1.3.1 Classical machine learning algorithms

Usually data is collected and converted into machine readable numerical features, which is a fixed length numerical matrix, and this is then used as an input for a model, which can be used for supervised, unsupervised classification or regression problem. Supervised learning is when the training data is labeled with class, for e.g. the data that corresponds to cancer and rest of the data that corresponds to healthy. This is known as binary class classification and if there are more than two classes then it is a multi-class classification. The best models for these types of classifications are SVMs, random forest (RF), decision trees and neural networks. Unsupervised learning is done when the training data are not labelled and the model tries to find a pattern in the dataset to cluster the data into groups. Methods like hierarchical clustering, k-nearest neighbor, k-means are example of unsupervised learning. Feature extraction and data normalization are the crucial steps before training the model. Features are numerical information that distinguish the classes and show patterns.

Support vector machines

SVMs[87], since 1996 have become the most widely used classical machine learning algorithm for linear and non-linear supervised learning classification. SVM is mostly used

for binary classification and can be extended to multiclass classification. It can also do regression analysis and here the algorithm is known as Support vector regressor[88]. To classify simple linear data into two classes SVM uses a hyperplane to divide it in a high dimensional space. The hyperplane is defined by a weight vectors w and an intercept b shown in equation 1.1, where x denotes the sets of features.

$$D(x) = w.x + b \tag{1.1}$$

As one can divide the group with many different hyperplanes, the task is to maximize the distance between the hyperplane and the nearest training data point known as margin maximization and margin is equal to the $2/|w|$. The features corresponding to data points on the margins are known as support vectors, which are used in the prediction of the unlabelled class. The parameter C, is to find the margin size which controls how many points can be misclassified. The soft margin increases the classifier generalizabilty. The size of soft margin are penalized by parameter C. Large value of C, corresponding to large penalties for misclassifcation and resembles a hard margin classifier and gamma measures the degree of misclassification. (see Figure 1.6).

$$\min_{\mathbf{w},b,\gamma,\xi} \ -\gamma + C\sum_{i=1}^{\ell}\xi_i \tag{1.2}$$

$$\text{subject to } y_i(\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle - b) \geq \gamma - \xi_i,\ \xi_i \geq 0 \tag{1.3}$$

$$i = 1,\ldots,\ell \ \text{ and } \ \|\mathbf{w}\|^2 = 1. \tag{1.4}$$

For non linear data SVM uses kernel functions to map the original finite space to higher dimension feature space by computing the inner products between the images of all pairs of data in the feature space. Most used kernel functions are:

$$\text{linear:} \qquad K(x_i, x_j) = x_i^T, x_j \tag{1.5}$$

$$\text{sigmoid:} \qquad K(x_i, x_j) = tanh(\gamma x_i^T, x_j + r) \tag{1.6}$$

$$\text{radial basis:} \qquad K(x_i, x_j) = exp(-\gamma|x_i - x_j|^2), \gamma > 0 \tag{1.7}$$

$$\text{polynomial:} \qquad K(x_i, x_j) = (\gamma x_i^T, x_j + r)^d, \gamma > 0 \tag{1.8}$$

where γ is the slope, d is the degree of the polynomial and r is a constant. The kernel trick is the $\mathbf{x}\cdot\mathbf{y} + \|\mathbf{x}\|^2\|\mathbf{y}\|^2$ which separates the feature in higher dimensional makes it possible to create a hyperplane in non linear datasets (see Figure 1.7).

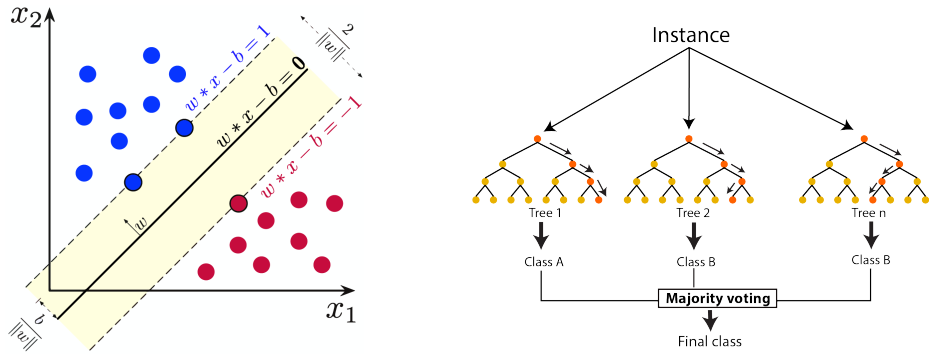


Figure 1.6: Support vector machines and Random Forests. (a) SVM a supervised classification learning algorithm classifies different classes by maximizing the hyperplane. Data points on the margin are called the support vectors. (b) RF have many decision trees predicting a class that forms a random forest and the majority vote defines the predicted class. Image source: Adapted from Wikipedia

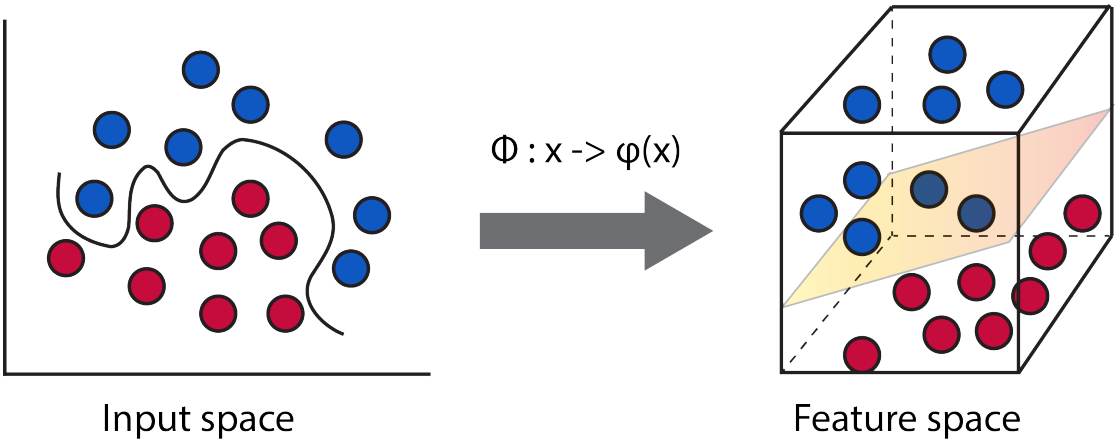


Figure 1.7: Kernel trick. Adapted from Wikipedia

To estimate the accuracy of prediction of the model cross-validation method is used. Dataset on which the model is trained known as training dataset, and a unknown data on which model is tested is known as validation or test dataset. Cross-validation is used to test the model's ability to predict new data that was not used in training the model, this helps to solve the problems like overfitting or selection bias and to give an insight on how the model will perform on an unknown dataset. One iteration of cross-validation involves partitioning the data into two subset, training the model on one subset (training set) and validating a model on another set (validation or test set). To reduce variability, multiple iteration of cross-validation are performed randomly selecting the percentage of data as two set, and the validation results averaged over the multiple iterations to give an estimate of the model's predictive performance.SVM model can be created using scikit, a python library, or in R and weka. SVM is also implemented in the Perseus software[18] and for example it can be used to find biomarkers using proteomics datasets. It has been successfully used to predict the subtype of breast cancer[89] and also for the prediction of subcellular localization with the dynamic organellar maps method[90,91].

Random Forest

Random forest[92] is one of the widely used supervised classification algorithm. They are ensemble learning method and can be used for both classification and regression models. Random forest contains large number of decision trees. It uses different features to create decision tree and outputs the class label in classification study and mean prediction of individual trees in regression study. Random forests was first proposed by Ho in 1995. The method to build forest of uncorrelated trees in CART, along with randomized node optimization and bagging was later described by Breiman[92] . RF uses bootstrap aggregating, or bagging. Given a training set $X = X_1, \ldots, X_n$ with labels $Y = Y_1, \ldots, Y_n$ bagging repeatedly samples a training set with a replacement and fits tree to these samples. for $b = 1, \ldots, B$

1. X_b, Y_b are sampled with replacement, n training example for X, Y.

2. Train a tree f_b on X_b, Y_b

The prediction for unknown samples are calculated by averaging the predictions from the training set from individual regression trees on $x^{'}$.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} -f_b(x^{'}) \tag{1.9}$$

or by taking majority votes in classification trees. The advantage of bootstrapping is that it decreases the variance of the model without increasing the bias. For uncorrelated trees, the prediction of a single tree is highly sensitive to noise but the average of many trees is not. Training many trees on a single training set will give strongly correlated trees or even same one many times. Bootstrapping gives uncorrelated trees by sampling different training set at each iteration. The number of trees B is a parameter, which needs to be tuned depending on size and nature of the training set. The optimum number of trees can be found using cross validation or by out-of-bag-error. In out-of-bag-error, mean prediction error on each training sample x_i is calculated. In RF along with bagging, random selection of subset of features are also included. If the features are very strong predictor of target output these features will be selected in many B trees which will result in making the trees correlated. The number of p features is also the parameter that needs to be tuned depending on the training dataset. The Figure 1.6 shows the random forest as ensemble of decision trees. Random forest and decision trees were used earlier to predict the MS/MS spectra intensities[68,93].

1.3.2 Neural networks

Neural networks or artificial neural networks have been through various phases. First mathematical model for neural network was developed in 1942 by McCulloch and Pits[94] then in 1949 pyscholoigist Hebb introduced first learning rule by memorizing and adapting the weights. Rosenblatt in 1958 introduced perceptrons. 1969 Minsky and Papert prove limitations of perceptron: one layer cannot represent even an XOR function. Then there was 13 year hibernation period in field of artificial neural networks. The second wave of research started with self organizing maps described by Kohonen in 1982. Since 1995, SVMstarted performing better than perceptrons and was widely used in various field along with RFand hidden Markov model (HMM). Third wave of neural network started from early 2000s and were named as deep learning as it has more than one hidden layer, In 2006, Hinton published work on pretraining of

multilayer neural network and Boltzmann machine. LeCun and Bengio developed convolutional neural networks. Recurrent neural network also improved and overcame all the limitation of vanishing gradient phenomenon by introducing LSTM. This could be possible because of the increased computational power and Graphical processing units (GPU) and huge amount of dataset available.

In this thesis, we will discuss deep feedforward neural networks and recurrent neural networks. These algorithms showed promising results in *article 1* and *article 2*.

Feed forward neural networks

Feed forward networks or multilayer perceptrons (MLPs) aim to approximate function f^*. Neural network are conceptualized based on biological neural networks. Mapping $y = f(x;\theta)$, where the feedforward network learns the value of the parameter θ that gives best function approximation. In Feedforward models the information flow through the function is being evaluated from input x, optimizing parameters and defining f, and finally getting the output y. There are no feedback connections where output of models are fed back to itself and when it has feedback connection it is known as recurrent neural networks, which is discussed in details in the next section. The fully connected feedforward neural networks are directed acyclic graph. The functions $f^{(1)}$,$f^{(2)}$,$f^{(3)}$are connected in chain to form function $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$[95]. Each input unit is connected to hidden layer and that is connected to the output layer, the final layer of the network. The length of chain gives depth of the network and this is called feed forward network as shown in figure 1.8. A loss function like mean squared error is used calculate the difference between true and predicted value. The architecture of neural network is the overall structure of the network: how many units it should have and how these units should be connected. A network with even one hidden layer is sufficient to fit the training set. A functional unit is also known as neuron because it is based on human brain structure. In a hidden layer, the functional unit is called hidden unit. It takes a vector as input and compute transformation z, and then applies element wise non-linear function $a(z)$. where z is:

$$z = W^T x + b \tag{1.10}$$

W is the weight matrix and b is the bias vector. They are the parameter associated with hidden layer. Parameters W and b are randomly initialized. They can be initialized as

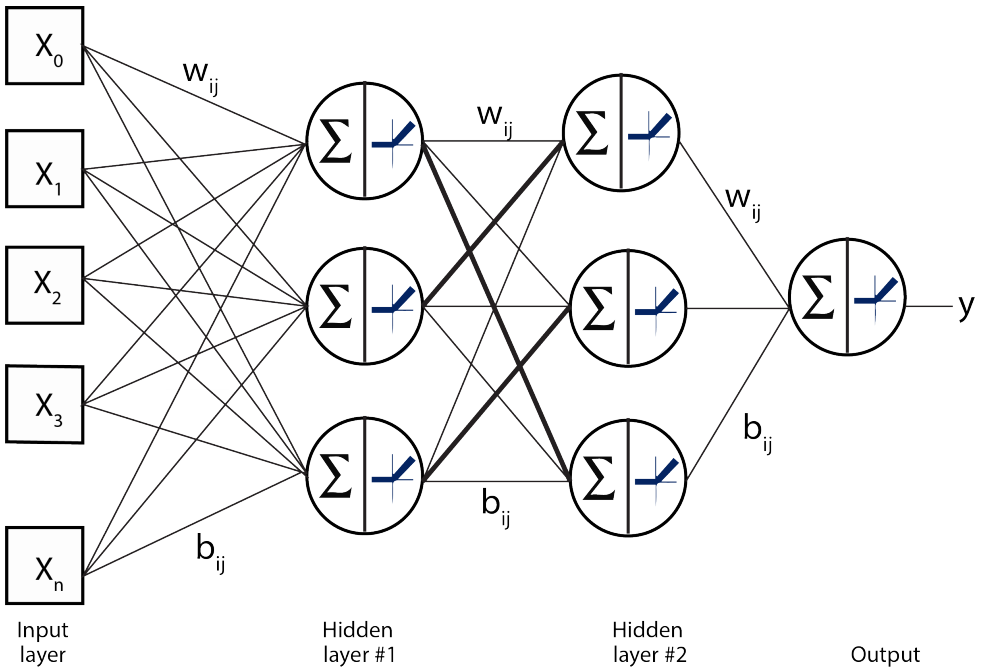


Figure 1.8: Fully connected feed forward neural network. It contains input layer, hidden layer and output layer. Hidden layer consists of hidden units which has two compartments, first computes the transformation z and then applies element wise non-linear function $a(z)$ and output of $a(z)$ is passed to next layer as input and output of last layer is the predicted y

zero or as random numbers.

$$W^l \sim \mathcal{N}(\mu = 0, \sigma^2 = \frac{1}{n^{[l-1]}}) \tag{1.11}$$

It states that weight matrix W in a particular layer l are randomly chosen from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 =$ multiplicative inverse of the number of neurons in layer $l - 1$. The bias b of all layers is initialized with 0.

$$a = \sigma(z) \tag{1.12}$$

where, a is an activation function using e.g. Sigmoid function. The neural networks are organized in groups of units called layers, with each layer being a function of the layer that preceded it.

$$z^{[1]} = (W^{[1]T}x + b^{[1]})a^{[1]} = \sigma(z^{[1]}) \tag{1.13}$$

and the second layer can be

$$z^{[2]} = (W^{[2]T}x + b^{[2]})a^{[2]} = \sigma(z^{[2]}) \tag{1.14}$$

and so on. The output of the last layer here $a^{[2]}$ gives the predicted $\hat{y}$. Then the loss function is calculated using mean squared error $\mathcal{L}(\hat{y}, y)$ to see how far the predicted $\hat{y}$ is from the true y.

Activation functions

The function f which determines the output of the neural networks is known as activation function. The function is attached to each neuron/hidden unit in the network, and determines whether it should be activated ("fired") or not, based on whether each neuron's input is relevant for the model's prediction. The activation function outputs a value in a range between 0 and 1 and between -1 and 1. User can choose different kind of non linear functions such as Sigmoid, TanH / Hyperbolic Tangent, Rectified Linear unit[96]. The sigmoid function is most widely used activation function. The output range of Sigmoid is from 0 to 1 (see Figure 1.9). Logistic function is defined:

$$logistic(x) = \frac{1}{1 + exp^{-x}} \tag{1.15}$$

For example, the rectified linear function $f(x) = max(0, x)$ is not differentiable at $x = 0$. The gradient descent preforms well even if Relu doesn't differentiate completely,

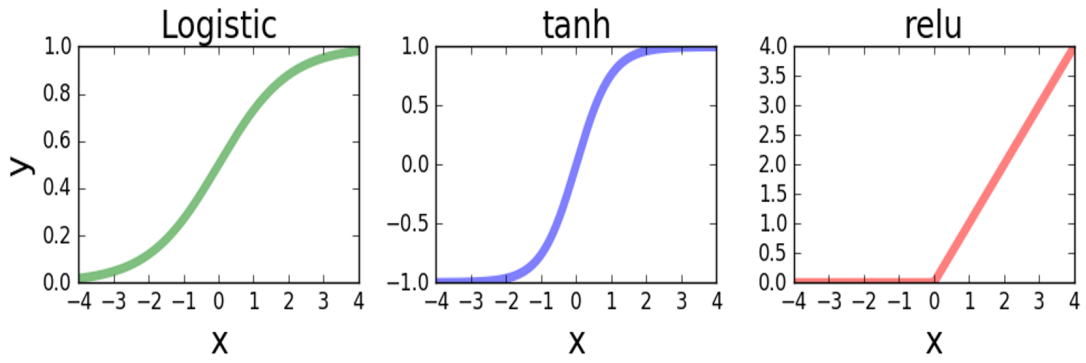


Figure 1.9: Activation functions. Logisitic, tanh, relu are the most widely used non linear activation functions in deep learning. Source wikipedia.

gradient descent performs well because we do not expect the function to reach the point where the gradient is 0. The rectified linear unit function is defined as following

$$f(x) = x^+ = \max(0, x) \tag{1.16}$$

$$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})} \tag{1.17}$$

Optimizers

Gradient descent is a way to minimize an objective function $J(\theta)$ parameterized by a model's parameters $\theta \in \mathbb{R}^d$ by updating the parameters in the opposite direction of the gradient of the objective function $\nabla_\theta J(\theta)$ w.r.t. to the parameters. The learning rate η determines the size of the steps we take to reach a (local) minimum. In other words, we follow the direction of the slope of the surface created by the objective function downhill until we reach a valley (see Figure 1.10). The gradiet descent method has few variables depending on the amount of data we take for the parameter optimization. In Stochastic gradient descent (SGD) each training point pair x^i, y^i is taken separately for the parameter updates.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)}) \tag{1.18}$$

Batch gradient descent method computes the gradient of the cost function w.r.t. to the parameters θ for the entire training dataset.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{1.19}$$

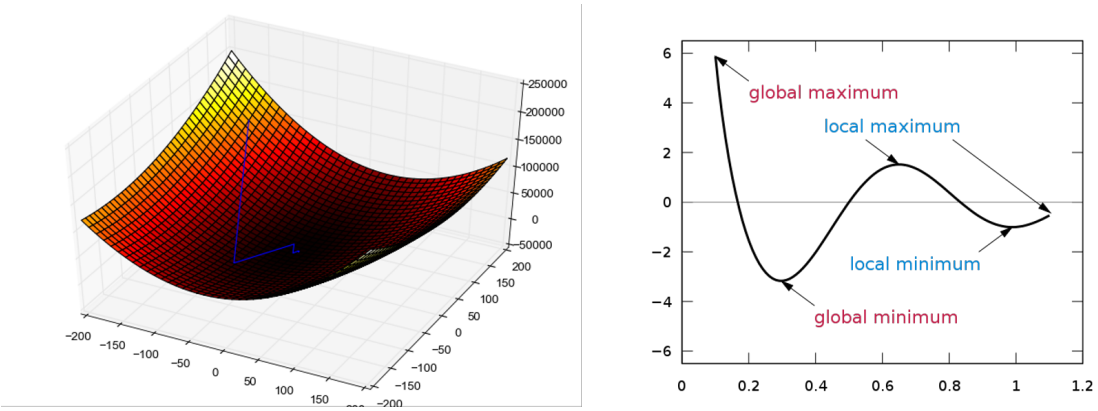


Figure 1.10: Stochastic gradient descent method. Source: Wikipedia.

Because we need to run the whole dataset through the deep learning model to calculate the gradients and perform just one update, batch gradient descent can be very slow and is intractable for datasets that do not fit in memory. While batch gradient descent converges to the minimum of the basin, where the parameters are placed in. SGD fluctuation enables it to jump to new and potentially better local minima. But it complicates convergence to the exact minimum, as SGD will keep overshooting. However, by slowly decreasing the learning rate, SGD can show same convergence as batch gradient descent, almost certainly converging to a local or the global minimum for non-convex and convex optimization respectively. **Mini-batch gradient descent** instead of taking complete training dataset it takes mini-batch of n training examples to performs an update.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)}) \tag{1.20}$$

This method reduces the variance of the parameter updates that can lead to more stable convergence and can make use of highly optimized matrix optimizations. It is used by all state-of-the-art deep learning algorithms. **Adagrad**[97] is able to deal sparse gradients and **RMSprop**[98] is able to deal with non-stationary object and it smooths the gradient. **Adam optimizer**[99], is a very computationally efficient gradient based optimization method for stochastic objective functions. It works on large datasets with high dimensional parameter spaces. It combines the advantages of AdaGrad and RMSprop (root mean squared prop)[98]. Adaptive Moment Estimation (Adam) is a robust and well-suited method for a wide range of non-convex optimization problems in the

machine learning field. It computes adaptive learning rates for each parameter.

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \end{aligned} \tag{1.21}$$

Dropout rate

Dropout[100] provides a computationally inexpensive powerful method of regularizing the models. It helps with over-fitting of model. The idea is to dropout (zero) randomly sampled hidden units and input features during each iteration of optimization. Adding dropout layer improved the ImageNet[85] classification which won several competitions. Alternatively, the procedure can be seen as averaging over many neural networks with shared weights.

Size of a mini-batch

For large training sets, it is suggested to divide the training data into small mini-batches of 10 to 100 cases before updating the weights. When the size of mini-batch changes it is important that learning rate should not change. Divide the total gradient computed on a mini-batch by the size of the mini-batch to avoid changes in the learning rate, so that learning rates are assumed to multiply the average, per-case gradient computed on a mini batch, not the total gradient for the mini batch. For training sets, first randomize the order of the training example and then mini-batch of size 10 can be used.Alternatively, the number of training cases should be divisible by mini-batch size. Choosing appropriate values of hyper parameters for new model applications requires heuristic learning ability. Ideal network architecture can be optimized by monitoring the validation set error and Google Vizier[101] provides one. Grid search is another way to do hyper-parameter tuning. Imagenet[85], an image classification model, that used a plethora of image data that is present on the internet, it uses a fully connected feedforward network together with convolutions neural networks and it performs even better than human in recognizing unknown images. Recurrent neural networks are used in language processing and machine translation problems. Now, with the user friendly API such as keras[102], which uses Tensorflow[103] as backend the implementation of feedforward network became much simpler for users of any field. This was also used **article 1** to implement wiNNer model to predict spectrum intensity.

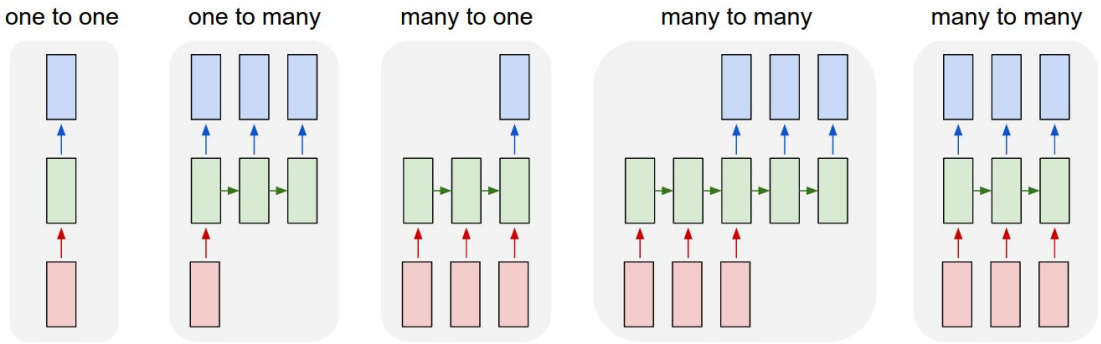


Figure 1.11: Types of sequence model based on input and output. Source: Andrej Karpathy blog.

1.3.3 Recurrent Neural Networks

Sequential data is kind of data where order matters like time series data or a sentence. So it's important to remember the past information for the prediction of correct output. Biological data (DNA, RNA and protein sequences) are sequential too. Recurrent neural networks allow operating over sequences of vectors: Sequences in the input, the output, or in both (see Figure 1.12). Depending on the size of input and output the models can be of many types (see Figure 1.11). The one-to-one model takes fixed-size input and fixed-size output (e.g.image classification). The one-to-many model takes fixed-size input but the output is a sequence of variable length (e.g. image captioning takes an image as input and outputs a sentence of words). The many-to-one model takes variable-length input and outputs the one-dimensional result (e.g. sentiment analysis where a given sentence is classified as a positive or negative sentiment). In the many-to-many model, both input and output are of variable length (e.g. Machine Translation: an RNN reads a sentence in English and then outputs a sentence in French). For biological sequence data, we need mostly many-to-one or many-to-many models (e.g. MS/MS spectrum prediction uses many to many, retention time prediction uses many to one model).

Recurrent models have a loop mechanism that is known as hidden state which is representation of previous inputs (see Figure 1.12). In contrary to the feed-forward networks that were discussed in the previous section, the input and output in recurrent neural networks (RNN) are recurrent. They can keep the information from the previous time points. The recurrent neural networks form a chain of repeating modules of a neu-

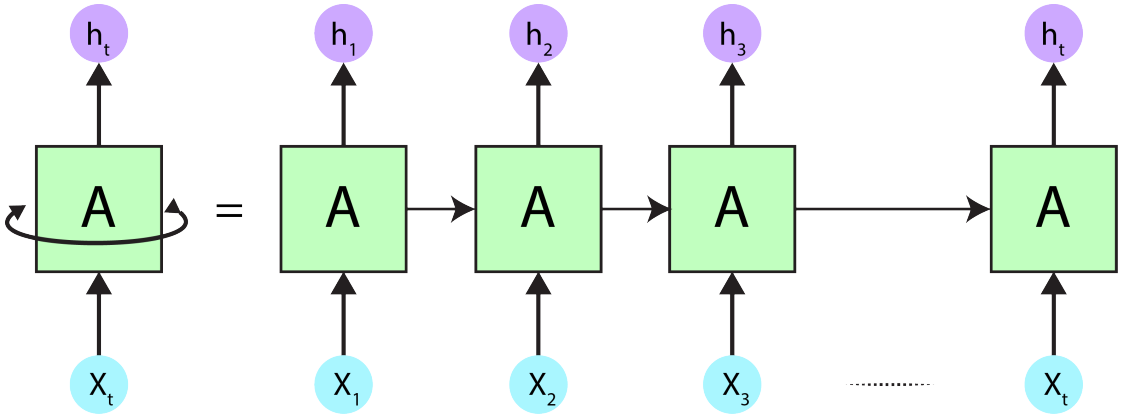


Figure 1.12: Unrolled recurrent neural networks at different time points t. Source: Adapted from colah's blog.

ral network. In standard RNN, the repeating module will have a very simple structure, with activation function such as a *tanh* layer. The limitation of RNN is vanishing gradient phenomenon. The vanishing gradient problem is when the gradient that is used to calculate the updated weights is vanishingly small, effectively preventing the weights from changing their value. Another problem is the exploding gradient that is an issue when the weights within a neural network have increased dramatically in magnitude in an unreasonable manner relative to their actual contribution to the model.

To overcome the limitations of vanishing gradient and to include long term dependencies of data at time point t, special RNN were developed by Hochreiter & Schmidhuber (1996)[104] known as LSTM. The LSTM layer contains blocks, that are called memory blocks. These blocks contains more than one cells and three gates (input, output and forget gate). The standard LSTM can be unidirectional where information is passed only in one direction or it could be bidirectional where input can be from past to future and reverse, which helps in capturing long term dependencies (see Figure 1.14). Similar to RNN, LSTM also have chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are three known as gates that interacts with each other. The gates have the output value of 0 and 1 and decide if the information is to be deleted or passed forward to the net gate.

The **Forget gate** f_t in equation 1.22 decides if the information should be kept or thrown away. The information from the previous hidden state h_{t-1} and current input state x_t is passed through forget gate, which outputs a number between 0 and 1 for each

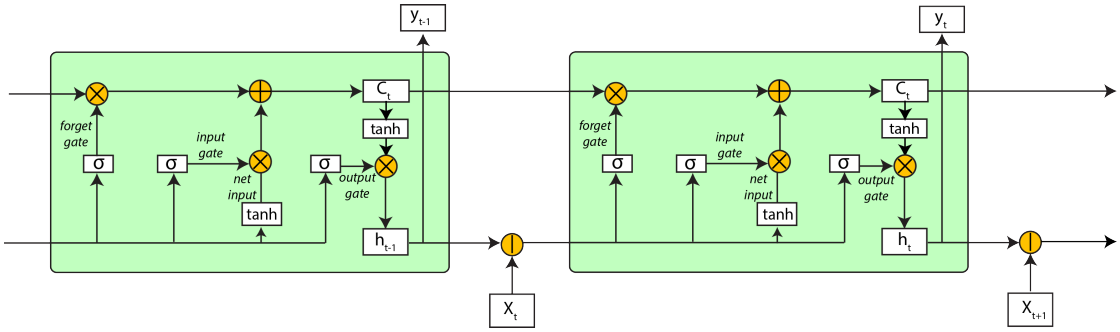


Figure 1.13: LSTM block containes four neural networks known as gates, forget gate, input gate and output gate. Adapted from Hermann Ney slides.

number in the cell state C_{t-1}. If the value is closer to zero is lost and if it is closer to 1 is kept.

Equations 1.23 and 1.24 decides which new information are going to be stored in the cell state C_t. First, **input gate** that decides which values will be updated and then, a *tanh* layer gives a vector, $\tilde{C}$, also known as net input that could be added to the state. Using equation 1.25 C_{t-1} is updated to C_t. Finally, the output is decided using the **output gate**. We will put cell state through *tanh* layer to get values between -1 and 1 and multiply it by the output of the sigmoid gate, which will be new hidden state (see figure 1.13). Varient of LSTM is gated recurrent unit (GRU) introduced by Cho, et al. (2014)[105]. It combines the forget and input gates into a single "update gate". It also merges the cell state and hidden state. The resulting model is simpler than standard LSTM models, and currently it is getting more popular.

$$\text{forget gate: } f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{1.22}$$

$$\text{input gate: } i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{1.23}$$

$$\tilde{C} = tanh(w_c[h_{t-1}, x_t] + b_c) \tag{1.24}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{1.25}$$

$$\text{output gate: } o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{1.26}$$

$$h_t = o_t * tanh(C^t) \tag{1.27}$$

σ -> represents sigmoid function

w_x -> weight for the respective gate(x) neurons

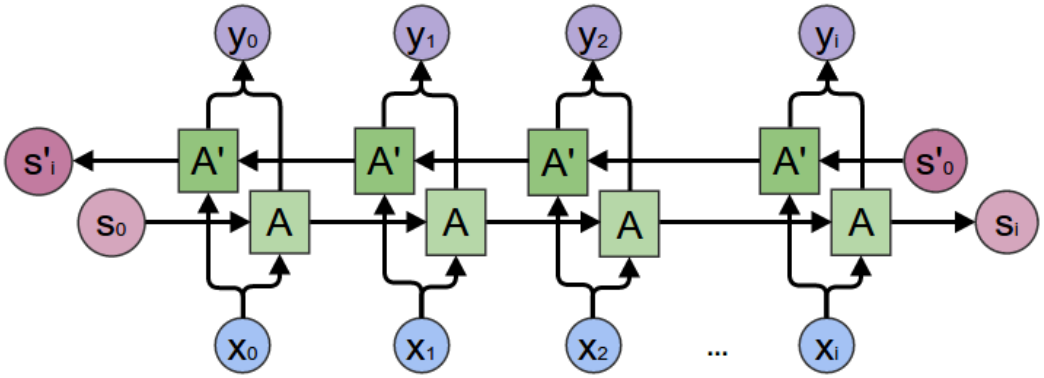


Figure 1.14: Bi-directional LSTM layer. Source: colah's blog.

h_{t-1} -> output of the previous lstm block(at timestamp t-1)
x_t -> input at current timestamp
b_x -> biases for the respective gates(x)
C_t -> cell state(memory) at timestamp(t).
$\tilde{C}_t$ -> represents candidate for cell state at timestamp(t) also known as net input
The LSTM architecture can be written using libraries such as Keras[102], Tensorflow[103], PyTorch and CNTK. Machine learning is always being used in the field of proteomics and protein sequence data analysis to find patterns, domains, biomarkers, and subcellular localization. Now with LSTM, we show in *article 1* that machine learning can learn the complex relationship between peptide sequence and peak intensities from MS and the predicted spectrum can now benefit in peptide identification.

1.4 Protein sequence features

Amino acids are the building block of proteins and they are arranged in the linear chain joined together by peptide bond. The polypeptide chain is the primary structure of the protein, which can be used to predict physicochemical, biological and functional properties of protein. So far, there are many sequence based predictor for various biological studies such as to predict signal peptides[106],transmembrane proteins[107], disorder regions[6], low complexity regions[7,108]. These predictors keep on being updated based on improvements in machine learning algorithms and dataset availability e.g. the first version of signal peptide prediction used neural networks[106], then used HMM[109] and

now it uses deep learning algorithm[110]. Protein and peptide sequences have also been used in secondary structure prediction[111] and to find domains and motifs (SMARTS[112], pfam[113]). Sequences can be used to learn patterns in two ways, one at the protein level taking complete protein sequence and other way is to use sliding windows approach, where a fixed size window slides over a few amino acid at a time. Using complete protein sequences, we can study protein crystallization[114]. Most common application of sliding windows approach is secondary structure predictions, signal peptides, transmembrane domain, and protein motifs prediction. The machine learning algorithms takes the numerical, fixed size input. So, we need to extract numerical feature from the protein and peptide sequences to give it as input to machine learning algorithm.

To create feature space from protein sequence, we can calculate frequency of each amino acid, dipeptide (two amino acids), tripeptide (three amino acids) in the protein sequence. The physico chemical properties (e.g. hydrophobic, hydrophilic, neutral residues) of amino acid sequence and its frequency can also be used as feature. Protein sequences contains various functional and structural regions, which can also be extracted as numerical features e.g. the number of disorder regions and the low complexity regions to predict functional properties of protein. Tools like PSIPRED (available as web based tool and standalone tool) predicts helical, sheet and coil residues in the protein sequence. These residues are used to calculate the frequency of helical, sheet and coil residues, which define the sequence attribute of secondary structure of the protein. Intrinsically unstructured/disordered regions are characterized by lack of stable secondary or tertiary structure under physiological conditions and in the absence of a binding partner/ligand. Disordered regions in proteins are predicted using DISOPRED[6]. DISOPRED predicts regions devoid of ordered regular secondary structure. SVM is used as a predictor model that takes the protein profile generated by psiblast as input. The important numerical features which can be used from these regions are the frequency of disordered residues, the length of disordered regions, the number of disordered regions, and the longest disordered region. The features were used in ***article 4*** to find out if the proteins of interest were enriched in the disorder regions and low complexity regions when compared to human proteome. The proteins that are rich in disorder regions and low complexity regions tend to aggregate when going through conformational stress. These aggregation cause neurodegenerative disorders such as Alzheimer's and Huntington's disease. Amino acid sequences can be also represented

as one-hot encoded vectors. One hot encoding is a binary representation, where one aa is on and rest 19 are off. Another method of amino acid representations is an embedding, where each aa is given a numerical value. These numerical representation of aa can be directly fed into machine learning-based models. In *article 1 and 2* we used one-hot encoding for peptide sequence representation.

1.5 MS/MS spectrum prediction

Database search is the most common strategy used in tandem mass spectrum identification. At a given retention time, the experimental spectra are matched to theoretical spectra, where information like mass and charge is considered but not the peak intensities. Peak intensity can be a piece of valuable information to correctly identify the spectra yet they are not fully exploited by mass spectra search engines[14–16,60,61].

It is not fully clear the physicochemical properties that effect the peptide fragmentation. Some of the factors, which effects the fragmentation are the following, precursor charge state, mass analyzers, the ionization methods. Peptide fragmentation method are very random because of the fluctuation in the ionization source and in the ion detection. Same peptides in the same experiment may produce different fragment spectra (known as technical replicates or baseline) and it varies more when different instrument, experimental designs or PSM algorithms are used. Theoretically, the intensities can be calculated from first principles by quantum chemistry. However, for molecules as large as peptides, this is too computationally expensive. Simpler models, such as the mobile proton hypothesis[115], exist for qualitative considerations, but they are not precise enough to be beneficial to the peptide identification process.

Intensities of fragment ion are varied based on residue on each side of target bond, and the types of ion formed (b-ions, y-ions etc) can give different intensities. The peak intensities cannot be quantified hence the database search engines don't used this information. This makes it ideal case for the prediction of MS/MS spectra intensities using machine learning algorithms. The goal of the prediction algorithm is to reach the limit of technical reproducible of the fragmentation spectra. Machine learning in proteomics is mostly used for preprocessing of the spectrum or post processing of the peptide identification.

In last decades, lot of research is done to predict fragment intensities using ma-

chine learning algorithms[69,93,116,117]. MassAnalyzer[118,119](used a kinetic model of peptide fragmentation), PeptideART[116](used neural networks) and MS2PIP (used RF). The input for these model were sequence features constructed manually. Feature space included amino acid residues in binary representation and compositional features as real values. The features that were used by most of the models are the amino acid composition of prefix and suffix aa residue of the target bond, length of both fragments, the first and the second neighbors of target bond, the parent mass, the ion masses, the N-terminal acetylation, gas-phase basicity, helicity, hydrophobicity, and the isoelectric point. The models were restricted by peptide length. Separate models were created for different charge and fragmentation methods(e.g. CID and HCD).

With the advancement of machine learning and deep learning algorithm it is now possible to give numerical representation of peptide sequence as input and the algorithm will learn the properties of fragmentation at each hidden layer. Bidirectional LSTM models show improvement in various fields like sentiment analysis, language translation, etc, and are proven to be the best algorithm choice to predict fragment spectra intensities. Bi-LSTM read the input sequence in both directions because fragmentation could be effected by both N- and C- terminus residues[120](see Figure 1.14). These predicted spectra intensities can be used in both DDA and DIA analysis.

Andromeda search engine[61] uses peak intensities only to give higher weights for matches to the high-intensity peaks. However, the PSM score calculation does not benefit from peak intensities. Peak intensities could be integrated in Andromeda scoring function in the following way: the Andromeda score of a given PSM can be replaced by a maximum of several attempts to score the same spectrum. One of these scores was the original Andromeda score. We can then calculate the score on subsets of the ions in theoretical spectra, always taking a certain number of top intensity peaks (by default it is top 3, 5, 7, 10 and 13 peaks) from the theoretical spectrum, which can include intensities predicted by machine learning algorithm.

In DIA mode, spectral library from DDA experiment is used to get the correct peptide identification and this can be both time and cost-effective. An *in-silico* generated spectrum library in DIA experiment will make DIA experiments much cheaper and faster and we can also overcome the problem of biased identification due to library taken from DDA experiments.

In ***article 1***, we aimed to develop two regression model to predict fragment in-

tensities with Pearson correlation coefficient (PCC) close to experimental reproducible, which we called as baseline. We trained bidirectional-LSTM algorithm using 160 million PSM, collected from PRIDE repository [4], on the GPU. This work was done in collaboration with Verily life Sciences. The model took 3 days to train and predicted intensities with PCC of 0.98 similar to baseline. As deep learning models are computationally expensive and needs huge dataset to train, we developed a simpler model based on sliding window approach which can be retrained easily with fewer data points whenever there are new datasets. We compared our model with the state-of-the-art predictor, MS2PIP [69] [68] and both the models performed better than Ms2PIP as shown in ***article 1***.

In ***article 1***, we presented proof-of-concept application of predicted intensities in both DDA and DIA acquisition mode. ***article 2***, shows that wiNNer model can easily be extended for new datasets. The predicted intensities were used to validate the ancient peptide spectrum matches.

Chapter 2

Manuscripts

2.1 High-quality MS/MS spectrum prediction

The state-of-the-art spectrum prediction methods discussed in the introduction section had few limitations such as models are not independent of peptide length, charge, fragmentation methods and types of mass analyzers. They also got sub-optimal prediction accuracy. To overcome these limitations we developed two regression models to predict spectra intensities. The first one termed DeepMass:Prism, deep learning-based model using bidirectional LSTM layer. Its predictive performance reaches the theoretical limit set by the reproducibility of technical replicates. The second approach, termed wiNNer (window-based neural network being easily retrainable), uses a sliding window-based machine learning method. The latter model is slightly inferior in accuracy to DeepMass:Prism. However, it is less computationally expensive to train, which makes ad hoc model creation for a given dataset more feasible. Both models can accommodate peptides of any length, unlike earlier approaches. In the deep learning approach, a single model can also accommodate multiple fragmentation types.

I was involved in collecting the training dataset from the PRIDE repository. We obtained 25 datasets from five different organisms(*Homo sapiens, Mus musculus, Escherichia coli, Saccharomyces cerevisiae, and Arabidopsis thaliana*) that contained 4,264 mass spectrometry runs. We processed the data using MaxQuant v.1.5.8.7. We obtained 60 million PSMs, 1.4 million unique sequence/charge state/fragmentation method/mass analyzer combinations.

I implemented the architecture of the wiNNer model, using Keras API with TensorFlow as a backend. We used python code to extract a sliding window-based feature matrix. The fixed length feature matrix was used as input to wiNNer architecture. DeepMass:Prism was developed by Verily Life Sciences group. To test its result and performance on CPU, I trained the model again and tested it. DeepMass:Prism predictor is accessible on google cloud and I was testing it continuously to make sure it is easily accessible to the users. We analyzed published HeLa whole-cell lysate data, obtained from Kelstrup et al[121] to compare the identification of MS/MS spectra using the conventional Andromeda scoring and the intensity-informed scoring. We predicted intensities for all candidate PSMs using DeepMass:Prism, wiNNer, and MS2PIP.

Later, I was responsible to integrate both the regression models in the MaxQuant environment. Both DeepMass:Prism and wiNNer used libraries written in python. To deploy the models in MaxQuant, we needed a C# libraries to use the saved DeepMass:Prism and wiNNer models for predictions. I used the TensorflowSharp library in C# for this.

Tiwary, Shivani, Roie Levy, Petra Gutenbrunner, Favio Salinas Soto, Krishnan K. Palaniappan, Laura Deming, Marc Berndl, Arthur Brant, Peter Cimermancic, and Jürgen Cox. High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, 16(6):519–525, June 2019. ISSN 1548-7105. URL https://doi.org/10.1038/s41592-019-0427-6

nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0427-6

High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis

Shivani Tiwary[1,5], Roie Levy[2,5], Petra Gutenbrunner[1,5], Favio Salinas Soto[1], Krishnan K. Palaniappan[2], Laura Deming[3], Marc Berndl[3], Arthur Brant[2], Peter Cimermancic[2*] and Jürgen Cox[1,4*]

Peptide fragmentation spectra are routinely predicted in the interpretation of mass-spectrometry-based proteomics data. However, the generation of fragment ions has not been understood well enough for scientists to estimate fragment ion intensities accurately. Here, we demonstrate that machine learning can predict peptide fragmentation patterns in mass spectrometers with accuracy within the uncertainty of measurement. Moreover, analysis of our models reveals that peptide fragmentation depends on long-range interactions within a peptide sequence. We illustrate the utility of our models by applying them to the analysis of both data-dependent and data-independent acquisition datasets. In the former case, we observe a *q*-value-dependent increase in the total number of peptide identifications. In the latter case, we confirm that the use of predicted tandem mass spectrometry spectra is nearly equivalent to the use of spectra from experimental libraries.

Peptide identification by fragmentation is a fundamental part of bottom-up mass-spectrometry-based proteomics[1,2]. Peptide molecules are fragmented with the aid of one of several technologies, including collision-induced dissociation[3] (CID), higher-energy collisional dissociation[4] (HCD) and electron transfer dissociation[5,6], producing a pattern of fragments that is indicative of the amino acid sequence[7]. The frequency with which a peptide backbone bond breaks determines the relative signal intensities in a fragmentation spectrum. Theoretically, the intensities can be calculated from first principles by quantum chemistry. However, for molecules as large as peptides, this is too computationally expensive to be practical. Simpler models, such as the mobile proton hypothesis[8], exist for qualitative considerations, but they are not precise enough to be beneficial to the peptide identification process. Hence, the intensity information contained in fragmentation spectra remains underused in many peptide identification strategies.

This problem is an ideal situation in which to employ machine learning. It can learn the relationship between sequence and fragment abundances based on a large dataset of training examples, without explicit knowledge of the physical mechanisms behind it. Furthermore, the predictive models do not have to remain black boxes, but can be examined with specialized methods that identify features or combinations thereof that are most relevant for making a prediction. While fragment intensity prediction has been attempted before using a variety of methods[9–12], they have had limited success. Here, we present a deep learning[13] method whose accuracy is close to the theoretical limitation. Furthermore, we demonstrate its utility by integrating it into data-dependent acquisition[14] (DDA) and data-independent acquisition[15] (DIA) computational proteomics workflows, and our results suggest that both can benefit from the improved spectrum prediction.

We developed two different regression strategies to model peak intensities. The first one, termed DeepMass:Prism, is a deep learning approach using a bidirectional recurrent neural network (RNN). Its predictive performance reaches the theoretical limit set by the reproducibility of technical replicates. The second approach, termed wiNNer (window-based neural network being easily retrainable), follows a classical sliding sequence window-based machine learning strategy. The latter model is less accurate than DeepMass:Prism. However, it has the advantage of being less computationally expensive to train, which makes ad hoc model creation for a given dataset more feasible. For both strategies, a single model can accommodate peptides of any length, unlike in other approaches[10]. In the deep learning approach, a single model can also accommodate multiple fragmentation types, or other dataset-specific information such as the fragmentation energy.

Results

Bidirectional RNN for spectral prediction. The problem of accurately predicting tandem mass spectra has long eluded conventional machine learning approaches for several reasons. First, because peptide sequences vary in length, they are incompatible with many algorithms that take fixed-length representation as an input. Second, different fragmentation and acquisition methods can be used to generate and acquire tandem mass spectra, each producing considerably different results. Moreover, precursor peptides are fragmented into different types of ions, where the abundance of any one ion type can be dependent on another. Training multiple models for each fragmentation method and ion type does not take advantage of such dependencies. Lastly, with large, publicly available mass spectrometry repositories, training of conventional algorithms (for example, support vector machines) becomes difficult, but complex, nonlinear approaches of deep learning become viable. Taking all of this into consideration, we selected RNNs[16]. RNNs are a class of artificial neural networks that are designed to work with sequential information, can accept inputs at different levels (for example, amino acid, peptide

[1]Computational Systems Biochemistry Research Group, Max Planck Institute of Biochemistry, Martinsried, Germany. [2]Verily Life Sciences, South San Francisco, CA, USA. [3]Google LLC, Mountain View, CA, USA. [4]Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway. [5]These authors contributed equally: Shivani Tiwary, Roie Levy, Petra Gutenbrunner. *e-mail: cpeter@verily.com; cox@biochem.mpg.de

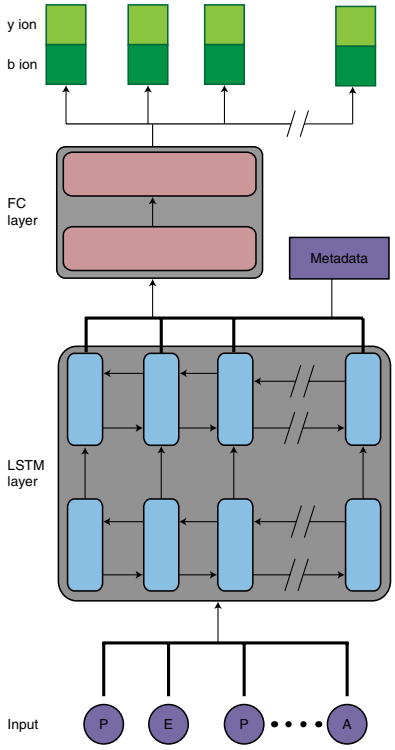


Fig. 1 | Bidirectional RNN architecture for the prediction of fragment intensities. The neural network contains two basic modules: an RNN encoder and a perceptron decoder. The encoder takes a one-hot-encoded peptide sequence as an input and outputs its fixed-length representation vector. The sequence representation vector is then combined with metadata features and input into the decoder. The decoder contains a set of fully connected layers that outputs intensities of different fragment ion types (for example, y and b ions) at each position in the input peptide sequence. FC, fully connected.

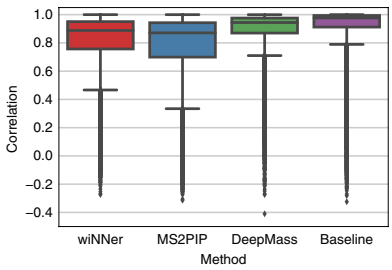


Fig. 2 | Comparing the performance of fragment ion intensity predictions for the different machine learning strategies. The box plots show distributions of PCCs between actual and predicted y- and b-ion peak intensities for each peptide in our testing dataset. The box plots contain 69,888, 69,888, 65,996 and 62,486 unique PSMs from the independent testing datasets for DeepMass:Prism, wiNNer, MS2PIP and technical variability, respectively. Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5 multiples past the interquartile range between the low and high quartiles. Data points beyond these ranges are considered outliers and are plotted as diamonds.

fragment and machine type) and of different types (for example, amino acid identities or their physicochemical properties), can predict multiple values simultaneously and can support training on datasets with millions of entries (Methods and Fig. 1).

Predictive performance of the bidirectional RNN. The datasets that are used for measuring the predictive performance are described in detail in the Methods. In brief, 25 complete datasets containing more than 60 million tandem mass spectrometry (MS/MS) spectra were used for training, testing and validation (Supplementary Fig. 1 and Supplementary Table 1). We first evaluated the performance of DeepMass:Prism against that of MS2PIP (ref. [10]) using the Pearson correlation coefficient (PCC) between true and predicted intensities for each peptide. When we compared all peptides in our testing set, we found that the accuracy of our model was markedly better than that of MS2PIP, with a PCC of 0.944 versus 0.871 (Fig. 2). We also calculated the PCC of repeatedly collected mass spectra of the same peptides to quantify technical variability in our dataset. Interestingly, our model's PCC nearly approached the theoretical maximum imposed by this measurement reproducibility of 0.976 (Fig. 2 and Supplementary Fig. 2).

DeepMass:Prism was highly accurate across CID and HCD fragmentation methods (median PCCs across all peptides in our testing dataset of 0.958 and 0.925, respectively; Supplementary Fig. 2) and for Fourier transform mass spectrometry and ion trap mass spectrometry mass analyzers (median PCCs of 0.924 and 0.949, respectively). The model was also accurate for precursor ion charges from 1 to 3 (median PCCs of 0.931, 0.952 and 0.901, respectively), with performance dropping at higher charges because of the lack of data. The length of the peptide only slightly affected the performance, with PCC for peptides with 5–10 amino acids being only marginally better than that for peptides with 30–35 residues (0.964 and 0.908, respectively). Similarly, the Andromeda score of a peptide-spectrum match (PSM) minimally affected the performance, with PCC for PSMs with a score of 200 being slightly lower than that for PSMs with a score of 700 (0.937 and 0.954, respectively). Finally, metadata features are crucial for accurate predictions; a model that does not take any metadata as inputs performs poorly (median PCC of 0.810; Supplementary Fig. 3).

Sliding-window-based prediction. Another way to construct a regression model that can be applied to peptides of variable sequence length is to use a sliding-window-based approach. Prediction of local protein properties on protein sequence windows has a long tradition and has been applied, for instance, to predict secondary structure, solvent accessibility and transmembrane regions[17,18]. The feature space is constructed from a sequence window centered on the backbone bond that is fragmented, extending k amino acids to the left and to the right from the backbone bond under consideration (Fig. 3). Amino acids at the N and C termini, their corresponding distance to the target bond and length of the peptide were also included in the feature space.

We applied three types of machine learning algorithms to this feature space: support vector machines[19,20], random forests[21] and a two-hidden-layer neural network. Of the three approaches, the neural network strategy wiNNer showed the best performance (Supplementary Fig. 4). Although wiNNer did not reach the prediction accuracy of DeepMass:Prism (Fig. 2), it performed better than the best existing conventional machine learning method for all tested combinations of precursor charge and fragmentation type. And, similar to DeepMass:Prism, with wiNNer only one model was needed to cover peptides of various lengths. Overall, this approach

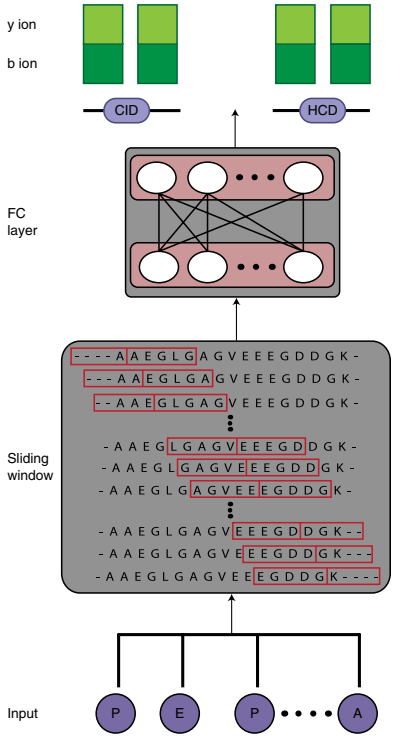


Fig. 3 | Sliding-window-based regression model for prediction of fragment intensities. A symmetrical sliding window is placed around the target peptide bond for which the b- and y-ion intensity should be predicted (red boxes). Amino acids in the window are translated into 0/1 variables by one-hot encoding. Additional features, including the amino acids at the N and C termini, the distances of the bond to the termini and the peptide length, are added to the feature space. This process is repeated for each position in the input peptide sequences. A fully connected two-hidden-layer neural network is then trained and outputs the logarithmic b- and y-ion intensities.

could be advantageous in situations where fast training is needed, such as when dataset-specific models are needed for the analysis of laboratory-specific proteomics data.

Interpretation of the DeepMass:Prism model. We first explored the agreement between the outputs of our RNN model and observed fragmentation efficiencies between different amino acid pairs[22] (Supplementary Fig. 5). For a set of A-A-A-[X]-[Z]-A-A-A-R peptides, where [X]-[Z] represents all possible combinations of amino acid residue pairs, we predicted the fragmentation efficiencies between the [X]-[Z] residue pair (Supplementary Fig. 5). Similar to the previous findings, our model reported notably higher fragmentation efficiency between [X]-Pro residues (where [X] can be any other residues), for both y- and b-ion types. The model also correctly predicted less efficient fragmentation between [X]-[Z] residue pairs where [X] is a hydrophobic residue. Furthermore, the model also correctly identified an increased fragmentation efficiency for b ions between His-[Z] pairs (Supplementary Fig. 5).

We next tested whether residues further up- or downstream from the site of fragmentation also contribute to the peak intensity assignment. Such long-range 'interactions' in peptide fragmentation have not been extensively studied, even though they can be observed in our dataset (Supplementary Fig. 6). For example, analysis of 63 pairs of peptides with a single residue mismatch showed notable differences in b-ion intensities 5–10 residues toward the C terminus from the mismatch site, while y-ion intensities showed less variability. Moreover, the fact that the window-based approaches that use small window sizes (Supplementary Fig. 7) performed poorly further supports the existence of these long-range interactions.

Based on these findings, we used our model to study long-range interactions. After randomly selecting 1,000 peptides from the independent testing set, we calculated the integrated gradients[23] over each position of the input peptide sequence and each ion type. Using these gradients, we attributed each peak's predicted intensity to the summed influence of every amino acid residue in the precursor fragment (Fig. 4); residues are able to positively or negatively influence any peak intensity. Finally, for each intensity prediction, we termed the most influential residues 'major attributions' (Methods), which can have both signs.

We found abundant evidence of long-range interactions, that is, major attributions not adjacent to the site of cleavage (Fig. 4 and Supplementary Fig. 8). Specifically, within b ions, major positive attributions were abundant toward the N terminus from the target residue, up to 75% of the length of the precursor peptide. While major negative attributions were notably less common, we observed many negative attributions a similar distance toward the C terminus from the target residue. We observed a different pattern in y ions: positive and negative attributions were more even on either side of the target residue. Nonetheless, while positive attributions were broadly spread about the length of the fragment, negative attributions were more tightly concentrated at the cleavage site, with a smaller cluster near the C terminus (an analysis on a per-residue basis is described in the Methods section).

DIA spectrum matching to predicted spectra. An accurate prediction of MS/MS peptide spectra is expected to benefit areas in which reference spectral libraries are utilized (and generated) for data analysis, as is the case for DIA and selected/multiple reaction monitoring[24,25]. The most common approach to analyze DIA data requires using a spectral library to determine the peptide identity. Although some library-free methods exist, such as DIA-Umpire[26] and DirectDIA, they are typically less sensitive than library-based methods. Accordingly, this necessitates performing a series of DDA experiments to build a sample-specific reference library. Given the stochastic nature of DDA, a single liquid chromatography–tandem mass spectrometry run is usually incomplete. Instead, replicate runs, and even sample fractionation, are typically required, further increasing experimental costs. Here, we tested whether in silico–generated spectral libraries, created using DeepMass:Prism, could replace those that are produced experimentally.

To evaluate this strategy, we processed a pooled human plasma sample into peptides using the Biognosys Sample Preparation Pro Kit, producing four replicate samples. We collected DIA data in triplicate for each sample, as well as DDA data in duplicate for one sample. A sample-specific spectral library was generated from the DDA data with Proteome Discoverer (Methods), and the DIA data were processed using Spectronaut[27]. The spectral library contained 7,441 peptides, of which 5,248 (71.0%) were identified and quantified on average (Fig. 5 and Supplementary Fig. 9). We then used DeepMass:Prism to generate an in silico spectral library for the same set of peptides and used it in another Spectronaut search. Specifically, the peak intensities for fragment ions for each of the 7,441 peptides from the DDA library were replaced with values from DeepMass:Prism. However, retention time information was preserved from the DDA-generated spectral library. In an ideal scenario, this approach would identify the same number of peptides

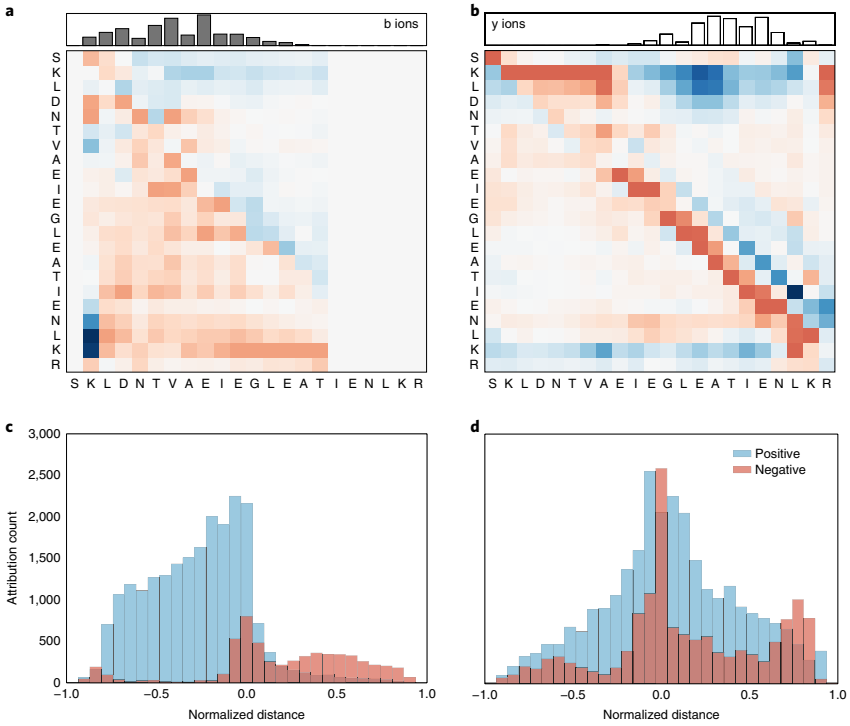


Fig. 4 | Interpretation of the DeepMass:Prism model. Left, y ions; right, b ions. Top: to help illustrate how the DeepMass:Prism prediction model works, the output for an example peptide sequence, SKLDNTVAEIEGLEATIENLKR, is plotted in an attribution table, which correlates the relationship between a position along the peptide sequence with the observed fragment ions. At the top are the scaled predicted peak intensities of the fragment ions corresponding to cleavage of the target residue at that position. Each pixel (*i*, *j*) in a heat map corresponds to the influence of residue *i* on peak intensity *j*. Blue pixels correspond to positive influence, and red to negative; values across each column *j* are normalized by the maximum value in that column. Bottom: histogram plots illustrate the distribution of distance between peak intensity and major attributions. Major attributions are attribution values with absolute values greater than or equal to 0.7 (based on heat map pixels in the top panel). Directional distances are normalized by peptide length; positive distances indicate a residue influencing cleavage toward the C terminus. Similar plots, but with unnormalized distances, are shown in Supplementary Fig. 8.

as when using the experimental generated library, and we came close—we identified and quantified on average 5,181 peptides, only 103 (1.9%) peptides fewer (Fig. 5), and with a high overlap (5,131 peptides or 97.1%). We repeated the analysis with a model that also predicted peptide retention time information in silico, with similar results (Methods).

Interestingly, the 103 peptides unidentified by DeepMass:Prism typically had low-confidence Spectronaut scores, with 46 peptides (45%) having q values worse than 10^{-3} in the DDA spectral library searches (Supplementary Fig. 9). Importantly, we observed highly correlated peptide abundance quantification in Spectronaut searches between the experimental library and the in silico library (PCC of 0.99; Supplementary Fig. 9c). As a control, we also generated a spectral library by predicting fragment intensities using MS2PIP (and preserving retention time information from the DDA data). We observed a much lower peptide identification rate compared with that for DDA libraries (on average, 3,976 or 75.8% peptides were identified; PCC of 0.96). Some of the difference between the numbers of peptides identified by MS2PIP- versus DeepMass:Prism-based libraries can be attributed to MS2PIP's lack in predicting spectra for peptides with modifications. However, even after the removal of methionine oxidation, the DeepMass:Prism-based library identified 4,661 peptides (17.2% more than MS2PIP).

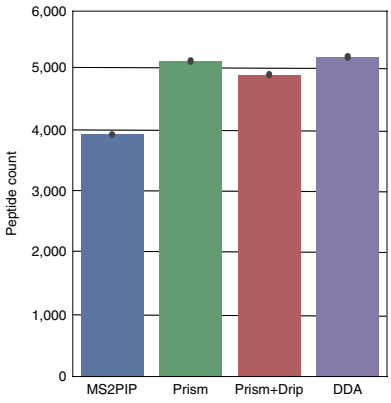


Fig. 5 | Application to spectral library generation for DIA data analysis. The bar plot compares the average number of identified and quantified peptides in plasma samples when using spectral libraries generated with MS2PIP, DeepMass:Prism, DeepMass:Prism including retention time prediction and the standard approach (that is, DDA experiments).

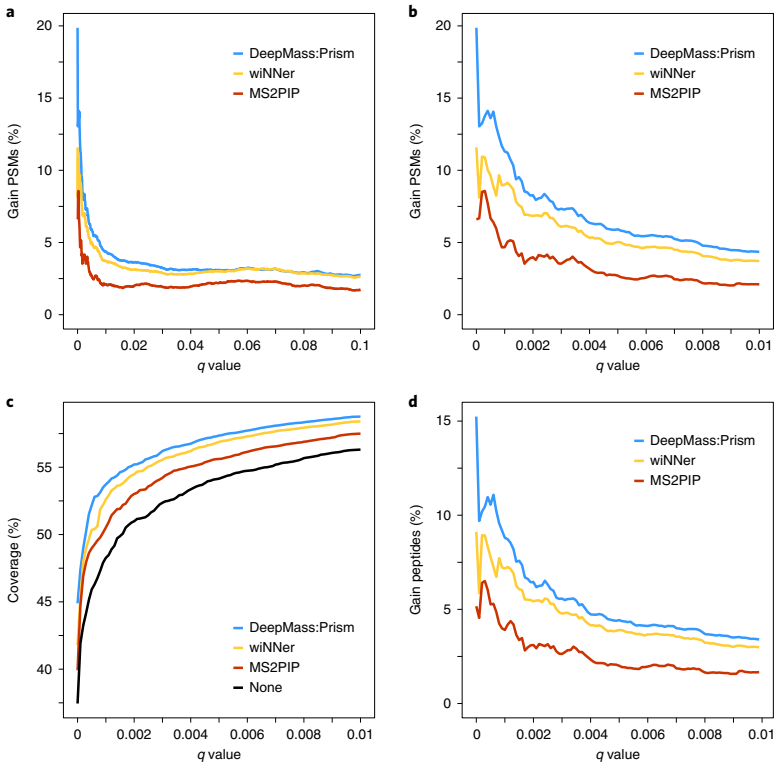


Fig. 6 | Application to PSM scoring for DDA data analysis. Identification-rate improvements on a complex HeLa dataset as a function of *q* value are shown for MS2PIP, wiNNer and DeepMass:Prism. The *q* values are calculated based on target-decoy search in MaxQuant. **a**, The percentage increase in PSMs compared with Andromeda searches without intensity prediction. The *q* value ranges from 0% to 10%. The curves are based on 146,606, 146,411 and 145,160 forward PSMs for DeepMass:Prism, wiNNer and MS2PIP, respectively. **b**, Same as **a**, but zoomed into the *q* value range from 0% to 1% and based on 127,806, 127,024 and 125,046 forward PSMs for DeepMass:Prism, wiNNer and MS2PIP, respectively. **c**, Coverage (that is, the percentage of identified MS/MS spectra) as a function of *q* value, without using intensity prediction and when using each of the three prediction models. **d**, The percentage increase in unique peptides compared with Andromeda searches without intensity predictions. The curves are based on 42,505, 42,335 and 41,793 peptides for DeepMass:Prism, wiNNer and MS2PIP, respectively.

Application to DDA peptide search engine scores. Intensity information in MS/MS spectra is highly informative and is expected to help in finding the correct PSM for a given MS/MS spectrum. Despite this, most search engines make no or little usage of the intensity information. For instance, the Andromeda search engine[28] uses peak intensities only to give higher weights for matches to the high-intensity peaks. However, it does not utilize expectations of peak intensities as they relate to sequence context when scoring PSMs. Here, we show that scoring PSMs can benefit from predicted fragment ion intensity information. We integrated intensity predictions into the Andromeda scoring function in the following way: the Andromeda score of a given PSM was replaced by a maximum of several attempts to score the same spectrum. One of these scores was the original Andromeda score. We then calculated the score on subsets of the ions in theoretical spectra, always taking a certain number of top intensity peaks (by default we took the top 3, 5, 7, 10 and 13 peaks) from the theoretical spectrum with intensities predicted by DeepMass:Prism. This strategy was similar to focusing on only the most intense transitions in the analysis of the selected reaction monitoring data. Naively, one may expect that reducing the number of theoretical peaks would reduce the number of matching fragments and hence reduce the score. Despite this, the Andromeda score of a match may still increase, even if the number of matches decreases, if the summed probability of finding this many or more matches by chance, given the number of provided theoretical fragments, decreases[28].

We compared the performances for a complex sample, in this case HeLa whole-cell lysate. Overall, we found that including intensity predictions increased both the PSM and peptide identification rates. We then compared the relative performance among DeepMass:Prism, wiNNer and MS2PIP by varying the PSM-level false discovery rate (FDR) in MaxQuant, which corresponds to scanning the *q* value (Fig. 5). Over the whole range of *q* values tested, there was a gain in PSM identifications when we used DeepMass:Prism and wiNNer, both of which outperformed MS2PIP (Fig. 6a). In particular, improvements through intensity prediction were the largest in the high-specificity range (that is, *q* value < 0.01) (Fig. 6b). Note that the improvements in MS/MS identification rates are on top of already high rates of ~50% for the conventional Andromeda search (Fig. 6c). In terms of gains in unique peptide sequences, DeepMass:Prism also outperformed wiNNer and MS2PIP (Fig. 6d). Finally, DeepMass:Prism also

outperformed wiNNer and MS2PIP in the number of identified protein groups (at 1% protein-level and PSM-level FDR), with a gain of 3.9% versus 2.3% and 2.1%, for wiNNer and MS2PIP, respectively (Fig. 5e).

Discussion

Through the use of machine learning, we demonstrate that MS/MS spectrum prediction can be nearly as accurate as the limits of technical reproducibility. Importantly, this can be taken advantage of in both DDA and DIA computational workflows to improve peptide identification rates and reduce the reliance on spectral libraries. Further, the deep learning regression models described here are highly interpretable, capturing how sequence features, including the interactions across multiple amino acid residues, contribute to peptide fragment ion abundance and the mobile proton hypothesis. Additionally, although the conventional window-based machine learning approach we described has slightly inferior predictive performance, it is less computationally intense to train. For both DDA and DIA application, integration of intensity prediction into the MaxQuant[29,30] environment is currently ongoing.

So far, we have restricted the spectrum predictions to peptides that are not carrying modifications, except for methionine oxidation. However, the generalization to PTMs is straightforward. The modified residues can be encoded as the 21st, 22nd and so on amino acids. Modification-specific neutral losses other than water and ammonia will need to be added for some modifications, as for instance serine and threonine phosphorylation. Non-tryptic peptides can already be accommodated with the current model, and predictions can currently being made for them. However, since the training dataset was from shotgun proteomics data submitted to the PRIDE database[31], it is biased toward tryptic peptides, and DeepMass:Prism performs better for these. As more data for non-tryptic peptides and more machine types become available, we will update our models to improve predictions for all peptides.

While these models to predict peptide fragment intensities will benefit peptide search engines, we anticipate that their greatest impact will be through providing in silico–based spectral libraries. For example, if an existing spectral library was generated using a fragmentation mode different from the data that needed to be analyzed, a new library could be generated in silico rather than experimentally. When it comes to the content of a spectral library, approaches are needed to construct a list of peptides based on previous knowledge. For example, if plasma samples are analyzed, proteins and peptides from the Plasma Proteome Database[32] could be used to construct the library. Another potential application will be to supplement an existing experimental library with a small number of hypothesis-driven peptides, such as peptides that harbor genetic variants, tumor mutations or post-translational modifications, expanding the range of interesting scientific and clinical questions beyond measuring the levels of proteins in a sample. A further option for analyzing cellular or tissue proteomes will be to construct an in silico library for the entire proteome. Such a spectral library could be supported by both RNA-sequencing data and peptide observability predictions[33,34]. We continue to explore these strategies and envisage a time when predicted spectral libraries will become a necessary and beneficial tool for proteomics and proteogenomics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0427-6.

Received: 22 June 2018; Accepted: 19 April 2019;
Published online: 27 May 2019

References

1. Cottrell, J. S. Protein identification using MS/MS data. *J. Proteom.* **74**, 1842–1851 (2011).
2. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational methods for understanding mass spectrometry-based shotgun proteomics data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).
3. Mitchell Wells, J. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzym.* **402**, 148–185 (2005).
4. Olsen, J. V. et al. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).
5. Coon, J. J., Syka, J., Shabanowitz, J. & Hunt, D. F. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques* **38**, 519–521 (2005).
6. Good, D. M., Wirtala, M., McAlister, G. C. & Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **6**, 1942–1951 (2007).
7. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
8. Boyd, R. & Somogyi, Á. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *J. Am. Soc. Mass Spectrom.* **21**, 1275–1278 (2010).
9. Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H. & Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput.* **230**, 219–230 (2006).
10. Degroeve, S., Martens, L. & Jurisica, I. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–3203 (2013).
11. Dong, N. P. et al. Prediction of peptide fragment ion mass spectra by data mining techniques. *Anal. Chem.* **86**, 7446–7454 (2014).
12. Park, J. et al. Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **14**, 909–914 (2017).
13. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
14. Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
15. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35–35 (2014).
16. Graves, A. et al. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 855–868 (2009).
17. Garnier, J., Gibrat, J.-F. & Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540–553 (1996).
18. Rost, B., Sander, C. & Schneider, R. PHD—an automatic mail server for protein secondary structure prediction. *Bioinformatics* **10**, 53–60 (1994).
19. Vapnik, V. N. *The Nature of Statistical Learning Theory* (Springer, 1995).
20. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **9**, 155–161 (1997).
21. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
22. Shao, C., Zhang, Y. & Sun, W. Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer. *J. Proteomics* **109**, 26–37 (2014).
23. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. Preprint at https://arxiv.org/abs/1703.01365 (2017).
24. Schubert, O. T. et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).
25. Wu, J. X. et al. SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Mol. Cell. Proteomics* **15**, 2501–2514 (2016).
26. Tsou, C.-C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
27. Bruderer, R. et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics* **16**, 2296–2309 (2017).
28. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
29. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
30. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
31. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).
32. Nanjappa, V. et al. Plasma proteome database as a resource for proteomics research: 2014 update. *Nucleic Acids Res.* **42**, D959–D965 (2014).
33. Mallick, P. et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).
34. Sanders, W. S., Bridges, S. M., McCarthy, F. M., Nanduri, B. & Burgess, S. C. Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics* **8**, S23 (2007).

Acknowledgements

This project has received funding from the European Union's EU Framework Program for Research and Innovation Horizon 2020 under grant agreement no. 686547 (S.T.) and from FP7 grant agreement no. GA ERC-2012-SyG_318987–ToPAG (J.C.). J.C. and P.G. are supported by the Marie Skłodowska-Curie European Training Network TEMPERA, a project funded by the European Union's EU Framework Program for Research and Innovation Horizon 2020 under grant agreement no. 722606. We thank E. Deutsch, J. Bingham, M. Liu, R. Perrone, P. Kheradpour, B. Brown, M. Edwards, L. Cao, N. Soltero, J. Lehar, T. Snyder, D. Glazer and T. Stanis for their help, support and suggestions.

Author contributions

S.T., R.L., P.G., F.S.S., P.C. and J.C. designed and developed the code, and performed the analyses. M.B. and L.D. helped with deep learning architecture design, as well as with reviewing the code and analyses. A.B. helped with data ingestion and preprocessing. K.K.P. carried out the wet-laboratory experiments and the DIA data analysis. R.L., P.C. and J.C. wrote the manuscript and directed the project. All authors read and approved the final manuscript.

Competing interests

R.L., A.B. and P.C. are employees of Verily Life Sciences. L.D. and M.B. are employees of Google LLC. Verily Life Sciences and Google LLC had no role in decisions related to the study/work, data collection or analysis of data described in this paper.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0427-6.

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

Correspondence and requests for materials should be addressed to P.C. or J.C.

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

Methods

Data. For evaluating the DeepMass:Prism model, we obtained 25 raw datasets from the PRIDE MS repository[31]. The data span five organisms (*Homo sapiens*, *Mus musculus*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*), and contain 4,624 liquid chromatography–mass spectrometry runs (Supplementary Table 1). The datasets were processed using MaxQuant v.1.5.8.7. The datasets comprise ~60 million MS/MS spectra in total (Supplementary Fig. 1), with 1.4 million unique sequence/charge state/fragmentation method/mass analyzer combinations. For each unique combination, a single representative MS/MS spectrum was used corresponding to the one with the best Andromeda score. Spectra with Andromeda scores below 100 were discarded. The unique combinations were randomly split into training, validation and testing sets with the ratio 90:5:5. The intensity values were normalized to a 0–10,000 range and were not log-transformed.

For evaluating the wiNNer model, we used the ProteomeTools dataset[35] (PXD004732). Different models were generated for CID + 2, CID + 3, HCD + 2 and HCD + 3, and for each one peptide sequence features were split into training, validation and testing sets with the ratio 90:5:5. All models contain unique peptides, which were selected taking the maximum Andromeda score; if this score was under 100, the peptide was discarded. Intensity values were normalized to 0–1 and then log-transformed using $\log_2(1 + \text{Intensity} \times 10{,}000)$.

Details of the RNN. Our model takes as an input a peptide amino acid sequence with its associated metadata, and returns intensities of different fragment ion types (that is, y and b ions with and without neutral losses) at each position along the input sequence (Fig. 1). The architecture of our neural network comprises two main modules: layers of recurrent cells (encoder) and layers of fully connected neurons (decoder). The encoder contains three bidirectional layers of long short-term memory[36] (LSTM) cells and emits a fixed-length representation of the input peptide. The fixed-length representation is concatenated to corresponding metadata (that is, precursor peptide length, charge state, fragmentation method and mass analyzer type) and then input into the decoder. Finally, the decoder outputs intensities of different fragment ion types at each position of the input sequence. Bidirectional LSTM cells and regular perceptron units with the rectifier activation function[37] were used to build the RNN and fully connected modules, respectively. The neural network was implemented in Tensorflow v.1.7.0 (ref. [38]). The learning and dropout rates, the number of layers, the number of hidden units in each module and the batch size hyper-parameters were optimized using Google Vizier[39] and the validation dataset. The model was trained on GPUs using the Adam optimization method[40]. The best model contained 3 bidirectional LSTM layers with 384 hidden units in each layer, and 4 fully connected layers with 768 neurons in each layer. Examples for the best and worst five predictions of the model can be found in Supplementary Fig. 10. A tab-separated text file with all spectrum predictions for the tryptic peptides in the human proteome (charge = 2, HCD) can be downloaded from the PRIDE dataset PXD010382 (uniprot-filtered-reviewed-human-peptides-ftms-hcd-charge2.tsv).

Details of the sliding-window-based machine learning. For the conventional machine learning approach on sequence windows, we use as feature space the adjacent amino acids around the backbone bond for which the y- and b-ion intensities should be predicted. In addition, the amino acids directly on the peptide N and C termini, as well as the distance of the bond to the termini and the length of the peptides, are used. Each amino acid feature was converted to 21 binary features by one-hot encoding. The 21st state represents cases in which the sliding window extends over a terminus of the peptide. The intensities of a single peptide were normalized by the maximum over y and b ions, and intensities for missing peaks were set to zero. For each fragment series, each backbone segment between two adjacent amino acids corresponds to an instance for the machine learning algorithm. Distances and peptide lengths were normalized so that they ranged from zero to one. In the wiNNer method we trained separate regression models for each combination of precursor charge and fragmentation type, and report results for CID + 2, CID + 3, HCD + 2 and HCD + 3.

To evaluate the wiNNer model, we calculated the PCC between true and predicted peak intensities for each peptide in the testing dataset. We used Keras (https://keras.io) v.2.0.8, a high-level neural network application programming interface, to train a simple two-layer neural network model. TensorFlow v.1.3.0 was used as backend in Keras. The architecture of the neural network includes two hidden layers. The model was trained for y and b ions in the same neural network model; hence there are two output units. The input layer contains 549 features for a window size of 24. We repeated this analysis for sequence window sizes of 4, 8, 16 and 24 residues, the result of which is shown for CID + 2 in Supplementary Fig. 7. The PCC increases monotonically with window size for all combinations of precursor charge and fragmentation type. Hence, we selected a window size of 24 for further analysis. Note that most peptides completely fit into the window since they are shorter than 24 residues. The number of hidden layers used for this window size is 312. Batch size, number of epochs, dropout and learning rate were optimized for all of the models separately. To test the window-based model, we used the independent test data taken from the 25 datasets, as shown in Supplementary Fig. 1. The PCCs of CID + 2, CID + 3, HCD + 2 and HCD + 3 models were 0.898, 0.793, 0.882 and 0.762, respectively. All four wiNNer models performed better than MS2PIP, as shown in Supplementary Fig. 11.

In addition to the neural network model, we explored support vector regression (SVR) and random forests from scikit-learn (http://scikit-learn.org, v.0.19.1) as machine learning algorithms. Because of practical limitations in training set size, we had to train the SVR on batches of 100,000 instances, the outputs of which were averaged. A radial basis function kernel was used, for which the width parameter was tuned in cross-validation. Then we trained a random forest in which we could use all training instances in one model. As a further option, we put a random forest layer on top of the output of SVR. Among all of these options, the neural network approach showed the best performance. For the comparison of different machine learning approaches, CID + 2 was used with a sliding window size of 8, which reduced the computing power needed as compared with the optimal window size of 24. This is why performance is lower here on average compared with the final model. However, we believe that relevant conclusions can be drawn for relative comparisons between machine learning methods.

Benchmarking and comparison with known methodologies. Using the validation and testing datasets, we compared the performance of our models with that of MS2PIP, taking into account only singly charged y- and b-ion series. Although DeepMass:Prism is capable of predicting peak intensities of fragments with neutral losses of water and ammonia, we had to base the comparisons on y- and b-ion peak intensities owing to the limitations of the other predictor. The MS2PIP server (https://iomics.ugent.be/ms2pip) with default settings was used for all analyses. We also compared the performances of our models with the best possible theoretical performance. We determined this by calculating the technical variability between spectra in our validation and testing sets against their random replicate observation in the entire dataset.

Plasma sample processing. EDTA plasma samples collected from three healthy patients were pooled together and then clarified by centrifugation at 17,000*g* for 10 min at 4 °C. Aliquots were prepared and stored at −80 °C. Immediately before processing, plasma aliquots were thawed at room temperature. Subsequently, four replicate samples were prepared following the Biognosys Sample Preparation Pro Kit. Each sample was transferred into a LoBind tube (Eppendorf), dried by vacuum centrifugation and then stored at −80 °C.

Mass spectrometry for plasma samples. Dried peptide samples were resuspended by the addition of 20 µl of 0.1% formic acid in water and water bath sonication for 10 min. Samples were subjected to centrifugation at 17,000*g* for 5 min at 4 °C. Subsequently, 18 µl was transferred into a new LoBind tube (Eppendorf) followed by the addition of 2 µl of 10x iRT (indexed retention time) solution (Biognosys). Liquid chromatography–tandem mass spectrometry experiments were performed using 1-µl injections. Samples were subjected to reversed-phase chromatography with an Easy-nLC 1000 HPLC (Thermo Scientific) connected in-line with a Q Exactive Plus (Thermo Scientific) mass spectrometer. External mass calibration was performed before analysis. A binary solvent system consisting of buffer A (0.1% formic acid in water (v/v)) and buffer B (0.1% formic acid in 95% acetonitrile (v/v)) was employed. The mass spectrometer was outfitted with a nanospray ionization source (Thermo Nanoflex source). The liquid chromatography was performed using a PepMap100 3-µm C18 (75 µm × 2 cm) trapping column followed by a PepMap RSLC 2-µm C18 (75 µm × 25 cm) analytical column. For both DDA and DIA experiments, the same 120-min biphasic method was used, consisting of a gradient from 4% to 25% buffer B for 105 min followed by 25% to 35% for 15 min, at a flow rate of 300 nl min^{-1}.

DDA of plasma samples. Each full-scan mass spectra was recorded in positive ion mode over the *m/z* scan range of 375–1,700 in profile mode at a resolution of 70,000. The automatic gain control (AGC) target was 3×10^6 with a maximum injection time of 50 ms. The 12 most intense peaks were selected for HCD fragmentation. Tandem spectra were collected at a resolution of 17,500 with an AGC target of 1×10^5 and maximum injection time of 60 ms. Dynamic exclusion and charge state screening were enabled, rejecting ions with an unknown or +1 charge state. An isolation window of 1.5 and normalized collision energy of 28 were used when triggering a fragmentation event.

DIA of plasma samples. Two scan groups were employed. First, using the selected-ion-monitoring scan group, we recorded a full-scan mass spectrum in positive ion mode over the *m/z* scan range of 400–1,200 in profile mode at a resolution of 70,000. The AGC target was 3×10^6 with a maximum injection time of 100 ms. Next, the DIA scan group was used to acquire 32 DIA segments of 15 Da each at a resolution of 35,000. The AGC target was 1×10^6 with a maximum injection time of 120 ms. An isolation window of 20 and normalized collision energy of 28 were used when triggering a fragmentation event. A global inclusion list was used to define each DIA segment.

Mass spectrometric data analysis of plasma samples. DDA data were processed with Proteome Discoverer (v.2.1), using Mascot as search algorithm. Fixed modifications included cysteine carbamidomethylation, and variable modifications included methionine oxidation and N-terminal acetylation. The files were searched against the human UniProt proteome database (downloaded 17 February 2016). DIA data were processed with Spectronaut (v.11, Biognosys) using default settings.

DeepMass:Prism model interpretation. To interpret our model, we applied the method of integrated gradients. Integrated gradients attributes the predicted output of a neural network to the set of input features (analogous to inspecting the product of the input features and coefficients in a linear model). For a given peptide sequence and precursor metadata, this method indicates the influence between each amino acid residue in the peptide sequence and the predicted intensity of each spectrum peak. Residues can either positively or negatively influence a peak's predicted intensity. Essentially, this provides a square $N \times N$ attribution matrix denoting the influence of residue i on peak j, where N is the peptide length (Supplementary Fig. 5). The diagonal elements of this matrix represent the degree to which the peak intensity is influenced by the identity of the residue at the cleavage site, while off-diagonal elements denote long-range interactions, where a peak's intensity is influenced by the identity of a distant residue. To focus on the influence of peptide identity, we held constant the precursor-level metadata and did not calculate gradients at the context. Specifically, for all peptides analyzed, we assumed a +2 charge state, fragmentation by HCD and Fourier transform mass spectrometry mass analyzer. The sum along a column of the attribution matrix equals the predicted intensity of the represented peak. Columns of this matrix were normalized by their maximum value such that the most positive attribution to peak intensity had value 1.0; other attributions were scored relative to this value. As peak values are non-negative, all columns have at least 1 element equal to 1.0, yet negative attributions can decrease below −1.0 in extreme cases.

To determine distances between interactions (Supplementary Fig. 6), we used an attribution threshold of ±0.7; any normalized value more extreme than this threshold was considered a major attribution. The directed distances between this major attribution and cleavage site were determined such that positive distances corresponded to instances where the attributed residue was situated downstream of bond cleavage (toward the C terminus). Finally, distances were normalized to the length of each particular peptide, such that a distance of ±1.0 corresponded to a full fragment ion length.

To determine the influence of specific amino acids on peptide fragmentation, we repeated the analysis on a per-residue basis (Supplementary Fig. 12). Specifically, for each amino acid, we calculated the distribution of distances of major attributions. Except in a few notable exceptions, amino acids did not greatly deviate from the general trend already described (Fig. 4). Nonetheless, we saw relevant clustering of per-residue profiles in the positive attribution of b ions. Broadly, we observed hydrophilic amino acids clustering distinctly from large hydrophobic amino acids. Proline is expected to show distinct behavior because of the well-known proline effect[41]. Indeed, it represents a notable outlier, and had substantially longer-reaching positive influence on predicted intensities up- and downstream (Supplementary Fig. 12, upper-right plot). Similarly, among negative attribution profiles, we identified two trends. First, branched-chain amino acids and proline had an influence relatively concentrated at the cleavage site, and second, they had a smoother distribution of influence downstream; asparagine was a notable outlier, with its strongest influence on peaks just upstream of it.

Evaluation of intensity prediction models in Andromeda scoring performance. We analyzed published HeLa whole-cell lysate data, obtained from Kelstrup et al.[42] (list of raw files: 20161213_NGHF_DBJ_SA_Exp3A_HeLa_1ug_60 min_15000_01.raw, 20161213_NGHF_DBJ_SA_Exp3A_HeLa_1ug_60 min_15000_02.raw, 20161213_NGHF_DBJ_SA_Exp3A_HeLa_1ug_60 min_15000_03.raw).

To compare the identification of MS/MS spectra using the conventional Andromeda scoring and the intensity-informed scoring, we predicted intensities for all candidate PSMs using DeepMass:Prism, wiNNer or MS2PIP. For the search configuration parameters, Trypsin/P was specified as enzyme, carbamidomethylation of cysteine was specified as a fixed modification and no variable modifications were selected. Protein FDR control was disabled to report results dependent on a q value on the PSM level. The examples in Supplementary Figs. 13 and 14 were taken from dataset PXD004732 in the PRIDE archive (ProteomeTools).

When comparing the Andromeda scores without and with using intensity prediction on a dataset consisting of synthetic peptides and, hence, with known ground truth[35], we saw several examples for which the correct PSM was not the highest-scoring one for that MS/MS spectrum when the conventional Andromeda score was used but became the highest-scoring PSM when the intensity-informed Andromeda score was used (Supplementary Figs. 13 and 14). As illustrated in Supplementary Figs. 13 and 14, examples where the highest-scoring PSM changed after including intensity information included cases where two adjacent amino acids were swapped and cases where a completely new peptide sequence became the top-scoring PSM.

Retention time prediction. We constructed an RNN model with a bidirectional LSTM layer with 40 hidden units, followed by another LSTM layer with 20 hidden units. The last output in the output sequence was then fed into 2 dense layers, the first with 20 hidden neurons and hyperbolic tangent activation function, and the last with 1 neuron and linear activation function. The input to the model is a one-hot-encoded sequence of amino acid residues, and the output is a predicted iRT value for the input peptide. We used the Adam optimizer and mean squared error as the loss function, and applied the dropout rate of 0.2 to each layer. The model was implemented using the Keras library and Tensorflow backend. The model was trained on data collected in-house from three different samples: human plasma, HeLa cell lysate and yeast cell lysate. The 69,680 peptide-charge pairs were split into training, validation and test set in a 75:20:5 ratio. It is worth noting that this model is not complete yet—it was built to assess the maximum possible peptide identification rate when a spectral library is completely generated in silico (that is, predicting both fragment ion intensities and retention times for each peptide sequence). As the model was trained using a limited number of samples that were subjected to liquid chromatography conditions identical to the DDA and DIA data obtained for plasma samples, we have not evaluated how well the model will generalize (that is, different liquid chromatography and column systems). It is even conceivable that the future model will have to be fine-tuned on each liquid chromatography–mass spectrometry setup independently.

To evaluate the full potential for generating spectral libraries completely by computation (that is, peptide fragment ion abundances and peptide retention time information are generated in silico), we predicted both precursor peptide retention times and MS/MS spectra. This combined model we termed DeepMass:Drip. It predicts iRT-calibrated retention times with high accuracy (R^2 of 0.97 and a median error of 4.84 as compared with R^2 of 0.88 and median error of 8.68 for SSRCalc[43]). To evaluate this method, we first generated a spectral library using DeepMass:Prism + Drip (that is, predicting fragment ion intensities and retention times for each of the 7,441 peptides from the DDA library). After performing Spectronaut searches, we quantified on average 4,957 peptides, 291 (5.5%) peptides fewer than the DDA library (Fig. 4). Here, too, peptides unidentified by DeepMass:Prism + Drip had low-confidence Spectronaut scores during the DDA library searches, with 118 (41%) peptides having q values worse than 10^{-3} (Supplementary Fig. 9).

Statistics. In box plots, each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5 multiples past the interquartile range between the low and high quartiles. Data points beyond these ranges are considered outliers and are plotted as individual data points. Numbers of data points used in each box plot are provided in the respective figure legends. The q values for PSM FDRs were estimated in MaxQuant with its standard target-decoy search strategy.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics data including summary tables have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD010382.

Code availability

We offer the trained DeepMass:Prism model for use via the Google Cloud ML platform (https://github.com/verilylifesciences/deepmass/tree/master/prism). To obtain the trained DeepMass:Prism model to run locally, please contact the corresponding authors. A user-friendly interface will be made available in the future MaxQuant releases.

References

35. Zolg, D. P. et al. Building proteometools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).
36. Hochreiter, S. & Schmidhuber, J. J. Long short-term memory. *Neural Comput.* **9**, 1–32 (1997).
37. Hahnioser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J. & Seung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **405**, 947–951 (2000).
38. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* 265–284 (USENIX Association, 2016).
39. Golovin, D. et al. Google Vizier: a service for black-box optimization. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1487–1495 (ACM, 2017).
40. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2015).
41. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl Acad. Sci. USA* **83**, 6233–6237 (1986).
42. Kelstrup, C. D. et al. Performance evaluation of the q exactive hf-x for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
43. Krokhin, O. V. Sequence-specific retention calculator. ALGORITHM for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-Å pore size C18 sorbents. *Anal. Chem.* **78**, 7785–7795 (2006).

2.2 The dental proteome of Homo antecessor

In the ***article 1***, we described that the sliding-window-based MS/MS spectrum intensity prediction model wiNNer is easily retrainable and can be adapted for new datasets. In this article, the authors collected dental enamel samples *Homo antecessor* from Atapuerca (Spain) and *Homo erectus* from Dmanisi (Georgia) and used them for mass-spectrometry based proteomics analysis. To validate the dental enamel peptide spectrum matches they used the wiNNer algorithm[122]. The results show that the wiNNer model retrained for randomly cleaved and heavily modified peptides provides a predictive performance similar to that of the wiNNer model trained on modern, trypsin-digested samples, assuring accurate sequence identification for the phylogenetically informative peptides. The median PCC between true intensity and predicted intensity was 0.76.

My contribution in this article was to retrain the wiNNer model for the prediction of the phylogenetically informative peptide sequences in the ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*). The dataset acquired to train the model was very small, making wiNNer an ideal method, as it performs better even with the smaller datasets. Ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*) were divided into two groups, the groups that contains phylogenetically informative peptide sequences and the group that does not. I prepared the training dataset by taking a subset of the phylogenetically non-informative peptides, and adding a previously published dataset of enamel proteomes from Dmanisi fauna to increase the size of the training dataset. The dataset only has HCD fragmentation, so I build two models. HCD+2 contains 5,555 unique modification-specific peptides, and HCD+3 contains 692 unique modification-specific peptides. For each unique modification-specific peptide, I took the spectrum with the highest Andromeda score. Spectra with an Andromeda score below 50 were discarded. The retained data for each model was split into 80:10:10 ratio for training, validation and test sets, respectively. Test data was kept for evaluating the wiNNer model by calculating the PCC between true and predicted intensities for each peptide. The training data has non-tryptic peptides. The samples were processed in MaxQuant with the following added variable modifications: Oxidation (M), Deamidation (NQ), Gln->pyro-Glu, Glu->pyro-Glu, Oxidation (P), Carbamidomethyl (C), Dioxidation(MW), Oxidation (W), His->GluOH (H), His->Asp (H), Arg Ornithine, Phospho (STY) and Phospho (S). The peptide sequences containing these variable mod-

ifications were taken as input for the model. Each amino acid residue and modified amino acid residues were converted to a unique 38 binary feature by one hot encoding. I trained two regression models, one for HCD+2 and one for HCD+3. The architecture of wiNNer model was slightly modified to train the ancient PSMs. The architecture of neural network includes 5 dense layers. The input layers contain 991 features for a window size of 24. The hidden unit is reduced from 600, 400, 200 to 50 in subsequent dense layers, and the output layer contains 2 units for y- and b-ion peak intensities. Hyper-parameters such as batch size, dropout, learning rate and number of epochs were optimized separately for different models. Instead of Adagrad, Adam optimizer was used for this model. The wiNNer model can be accessed on GitHub https://github.com/cox-labs/wiNNer.git. The results show that the PCC between true and predicted intensities for each peptide in test sets of HCD+2 and HCD+3 models were 0.85 and 0.81 respectively. These results are close to the wiNNer model for unmodified sequences, where the PCC is 0.88 for HCD+2 and 0.76 for HCD+3. This shows a perfect example that wiNNer can be easily retrained and can be used whenever there is new dataset available (for e.g. TMT modifications, different fragmentation methods) or when training set is very small.

Frido Welker, Jazmín Ramos-Madrigal, Petra Gutenbrunner, Meaghan Mackie, **Tiwary, Shivani**, Rosa Rakownikow Jersie-Christensen, Cristina Chiva, Marc R. Dickinson, Martin Kuhlwilm, Marc de Manuel, Pere Gelabert, María Martinón-Torres, Ann Margvelashvili, Juan Luis Arsuaga, Eudald Carbonell, Tomas Marques-Bonet, Kirsty Penkman, Eduard Sabidó, Jürgen Cox, Jesper V. Olsen, David Lordkipanidze, Fernando Racimo, Carles Lalueza-Fox, José María Bermúdez de Castro, Eske Willerslev, and Enrico Cappellini. The dental proteome of homo antecessor. *Nature*, 580(7802):235–238, April 2020. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-020-2153-8

Article

The dental proteome of *Homo antecessor*

https://doi.org/10.1038/s41586-020-2153-8

Received: 4 July 2019

Accepted: 21 January 2020

Published online: 1 April 2020

Check for updates

Frido Welker[1,22]✉, Jazmín Ramos-Madrigal[1,22], Petra Gutenbrunner[2,22], Meaghan Mackie[1,3], Shivani Tiwary[2], Rosa Rakownikow Jersie-Christensen[3], Cristina Chiva[4,5], Marc R. Dickinson[6], Martin Kuhlwilm[7], Marc de Manuel[7], Pere Gelabert[7], María Martinón-Torres[8,9], Ann Margvelashvili[10], Juan Luis Arsuaga[11,12], Eudald Carbonell[13,14], Tomas Marques-Bonet[4,7,15,16], Kirsty Penkman[6], Eduard Sabidó[4,5], Jürgen Cox[2], Jesper V. Olsen[3], David Lordkipanidze[10,17], Fernando Racimo[18], Carles Lalueza-Fox[7], José María Bermúdez de Castro[8,9]✉, Eske Willerslev[18,19,20,21]✉ & Enrico Cappellini[1]✉

The phylogenetic relationships between hominins of the Early Pleistocene epoch in Eurasia, such as *Homo antecessor*, and hominins that appear later in the fossil record during the Middle Pleistocene epoch, such as *Homo sapiens*, are highly debated[1–5]. For the oldest remains, the molecular study of these relationships is hindered by the degradation of ancient DNA. However, recent research has demonstrated that the analysis of ancient proteins can address this challenge[6–8]. Here we present the dental enamel proteomes of *H. antecessor* from Atapuerca (Spain)[9,10] and *Homo erectus* from Dmanisi (Georgia)[1], two key fossil assemblages that have a central role in models of Pleistocene hominin morphology, dispersal and divergence. We provide evidence that *H. antecessor* is a close sister lineage to subsequent Middle and Late Pleistocene hominins, including modern humans, Neanderthals and Denisovans. This placement implies that the modern-like face of *H. antecessor*—that is, similar to that of modern humans—may have a considerably deep ancestry in the genus *Homo*, and that the cranial morphology of Neanderthals represents a derived form. By recovering AMELY-specific peptide sequences, we also conclude that the *H. antecessor* molar fragment from Atapuerca that we analysed belonged to a male individual. Finally, these *H. antecessor* and *H. erectus* fossils preserve evidence of enamel proteome phosphorylation and proteolytic digestion that occurred in vivo during tooth formation. Our results provide important insights into the evolutionary relationships between *H. antecessor* and other hominin groups, and pave the way for future studies using enamel proteomes to investigate hominin biology across the existence of the genus *Homo*.

Since 1994, over 170 human fossil remains have been recovered from level TD6 of the Gran Dolina site of the Sierra de Atapuerca[10] (Burgos, Spain) (Extended Data Fig. 1, Supplementary Information). These fossils have been dated to the late Early Pleistocene epoch and exhibit a unique combination of cranial, mandibular and dental features[9,11]. To accommodate the variation observed in the human fossils from TD6, a new species of the genus *Homo*—*H. antecessor*—was proposed in 1997[9]. The relationship of this species to earlier or later hominins in Eurasia—such as the *H. erectus* specimens from Dmanisi or Neanderthals, Denisovans and modern humans, respectively—have been the subject of considerable debate[3,4,12,13]. These issues remain unresolved owing to the fragmentary nature of hominin fossils at other sites, and the failure to recover ancient DNA in Eurasia that dates to the Early, and most of the Middle, Pleistocene epoch.

By contrast, recent developments in the extraction and tandem mass-spectrometric analysis of ancient proteins have made it possible to retrieve phylogenetically informative protein sequences from Early Pleistocene contexts[6,8]. We therefore applied ancient protein analysis to a *H. antecessor* molar from sublevel TD6.2 of the Gran Dolina site of the Sierra de Atapuerca (specimen ATD6-92) (Extended Data Fig. 2a). This specimen, identified as an enamel fragment of a permanent lower left first or second molar, has been directly dated to 772–949 thousand

[1]Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Copenhagen, Denmark. [2]Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany. [3]The Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark. [4]Center for Genomic Regulation (CNAG-CRG), Barcelona Institute of Science and Technology, Barcelona, Spain. [5]Proteomics Unit, University Pompeu Fabra, Barcelona, Spain. [6]Department of Chemistry, University of York, York, UK. [7]Institute of Evolutionary Biology (UPF-CSIC), University Pompeu Fabra, Barcelona, Spain. [8]Centro Nacional de Investigación sobre la Evolución Humana (CENIEH), Burgos, Spain. [9]Anthropology Department, University College London, London, UK. [10]Georgian National Museum, Tbilisi, Georgia. [11]Centro Mixto UCM-ISCIII de Evolución y Comportamiento Humanos, Madrid, Spain. [12]Departamento de Paleontología, Facultad de Ciencias Geológicas, Universidad Complutense de Madrid, Madrid, Spain. [13]Departamento d'Història i Història de l'Art, Universidad Rovira i Virgili, Tarragona, Spain. [14]Institut Català de Paleoecologia Humana i Evolució Social (IPHES), Tarragona, Spain. [15]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain. [16]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. [17]Tbilisi State University, Tbilisi, Georgia. [18]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. [19]Department of Zoology, University of Cambridge, Cambridge, UK. [20]Wellcome Sanger Institute, Hinxton, UK. [21]Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark. [22]These authors contributed equally: Frido Welker, Jazmín Ramos-Madrigal, Petra Gutenbrunner.
✉e-mail: frido.welker@sund.ku.dk; josemaria.bermudezdecastro@cenieh.es; ewillerslev@sund.ku.dk; ecappellini@sund.ku.dk

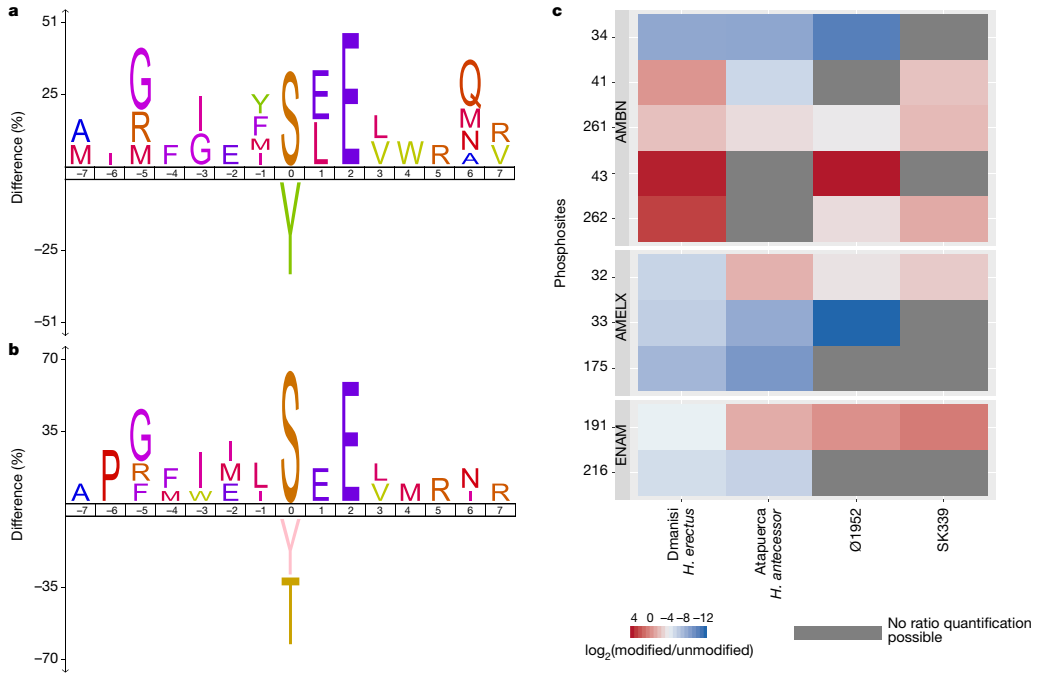


Fig. 1 | Phosphorylation of hominin enamel proteomes. a, Phosphorylation sequence motif analysis of *H. antecessor* specimen ATD6-92. **b**, Phosphorylation sequence motif analysis of *H. erectus* specimen D4163. **c**, Phosphorylation occupancy comparison, expressed as $\log_2$-transformed summed intensity ratio of modified and unmodified peptides, for amino acid sites for which data are available for at least two specimens. *y* axis labels indicate the position of the phosphorylated amino acids for each protein (UniProt accession numbers Q9NP70 (AMBN), Q99217 (AMELX) and Q9NRM1 (ENAM)). SK339 denotes an archaeological specimen from a modern human, which is approximately three centuries old (see 'Recent human control specimens' in the Methods for details).

years ago (ka) using a combination of electron spin resonance and U-series dating[11]. In addition, we sampled dentine and enamel from an isolated *H. erectus* upper first molar (specimen D4163) (Extended Data Fig. 2b) from Dmanisi (Georgia) that has been dated to 1.77 million years ago (Ma)[1,14,15], as amino acid racemization analysis of this specimen indicated the presence of an endogenous protein component in the intracrystalline enamel fraction of the tooth (Extended Data Fig. 3, Supplementary Information). On both specimens, we performed digestion-free peptide extraction optimized for the recovery of short, degraded protein remains[6]. Nanoscale liquid chromatography–tandem mass spectrometry (nanoLC–MS/MS) acquisition was replicated in two independent proteomic laboratories (Extended Data Table 1), implementing common precautions and analytical workflows to minimize protein contamination (Methods). We compared the proteomic datasets retrieved from the Pleistocene hominin tooth specimens with those generated from a positive control, a recent human premolar (Ø1952; which is from a male individual and is approximately three centuries old), as well as previously published Holocene teeth[16] (Methods, Supplementary Information). Finally, to validate our enamel peptide spectrum matches, we performed machine-learning-based MS/MS spectrum intensity prediction using the wiNNer algorithm[17]. The results show that the wiNNer model retrained for randomly cleaved and heavily modified peptides provides a predictive performance similar to that of the wiNNer model trained on modern, trypsin-digested samples, assuring accurate sequence identification for the phylogenetically informative peptides (median Pearson correlation coefficients of ≥0.76) (Methods, Supplementary Fig. 6, Supplementary Information).

Protein recovery from the Dmanisi dentine sample was limited to sporadic collagen type I fragments, and therefore in-depth analysis of this material was not further pursued. By contrast, we recovered ancient proteomes from both hominin enamel samples. We found that the composition of these proteomes is similar to that of the recent human specimen that we processed as a positive control, as well as to previously published proteomes from ancient enamel[6,16,18,19] (Extended Data Table 2, Supplementary Table 6). The enamel-specific proteins include amelogenin (both AMELX and AMELY isoforms), enamelin (ENAM), ameloblastin (AMBN), amelotin (AMTN) and the enamel-specific protease matrix metalloproteinase 20 (MMP20). Serum albumin (ALB) and collagens (COL1α1, COL1α2 and COL17α1) are also present. For the enamel-specific proteins, the peptide sequences that we retrieved cover approximately the same protein regions in all of the specimens that we analysed (Extended Data Fig. 4). Although destructive, our sampling of Pleistocene hominin teeth resulted in higher protein sequence coverage than acid-etching of Holocene enamel surfaces[16,20] (Supplementary Fig. 7). The AMTN-specific peptides largely derive from a single sequence region involved in hydroxyapatite precipitation through the presence of phosphorylated serines[21]. Finally, the observation of the AMELY-specific peptides (which is coded on the non-recombinant portion of the Y chromosome) demonstrates that the *H. antecessor* molar that we studied belonged to a male individual[16] (Extended Data Fig. 5).

Besides proteome composition and sequence coverage, several further lines of evidence independently support the endogenous origin of the hominin enamel proteomes. Unlike exogenous trypsin, keratins and other human-skin contaminants that we identified, the enamel

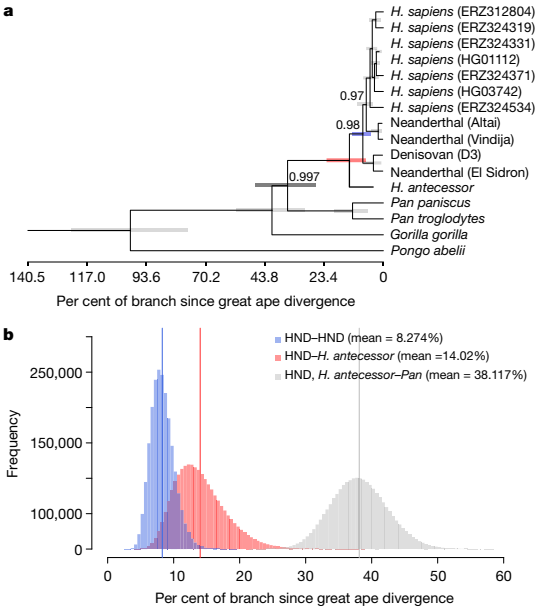


Fig. 2 | Phylogenetic analysis of *H. antecessor* ATD6-92. a, Maximum credibility tree estimated using BEAST and a concatenated alignment of seven protein sequences recovered for the ancient sample. Posterior Bayesian probabilities are indicated at nodes with a probability of ≤1. Horizontal error bars at each node indicate the 95% highest posterior density intervals for the split time estimates. The position of *H. antecessor* is consistent with that obtained via maximum likelihood (Supplementary Fig. 13) and Bayesian (Supplementary Fig. 16) analyses. ERZ and HG codes in parentheses after *H. sapiens* refer to identifiers for data from the Simons Genome Diversity Panel[33] and 1000 Genomes Project[34], respectively (see 'Comparison between the ancient protein sequences and modern reference proteins' in the Methods for details). **b**, Histograms of the divergence times obtained for the split between *H. antecessor* and the *H. sapiens*, Neanderthal and Denisovan clade (HND) (red), the HND–HND split (blue), and the *Pan*–(HND + *H. antecessor*) split (grey). Divergence times in **a** and **b** are shown as a percentage of the time since the divergence of all great apes.

proteins have high deamidation rates (Extended Data Fig. 6)—above the rate observed for the recent human specimens (Supplementary Fig. 8). Both Pleistocene hominins have average peptide lengths that are shorter than those observed for our recent human controls (Extended Data Fig. 6d). The average peptide length is shorter in the Dmanisi hominin, but longer in the younger Atapuerca hominin (Extended Data Fig. 6d). By contrast, we observe that the peptide lengths in enamel from the Dmanisi hominin are indistinguishable from those of the faunal remains from the same site. Together, our protein data are therefore consistent with theoretical and experimental[6,22] expectations for samples of their relative age.

In addition to diagenetic modifications, we observe two kinds of in vivo modification in our recent and ancient enamel proteomes. First, we detect serine (S) phosphorylation within the S-X-E motif (Fig. 1a, b). This motif, as well as the S-X-phosphorylated S motif, is recognized by the FAM20C secreted kinase, which is active in the phosphorylation of extracellular proteins[23,24]. The presence of phosphoserine in fossil enamel and its location in the S-X-E and/or S-X-phosphorylated S motifs has also previously been observed in other Pleistocene enamel proteomes[6,25]. Phosphorylation occupancy can be computed successfully for ancient and recent samples, and reveals differences in the ratios of phosphorylated peptides between samples (Fig. 1c, Supplementary Table 5). Second, the peptide populations that we retrieve primarily cover the ameloblastin, enamelin and amelogenin sequence regions, representing cleavage products deriving from in vivo activity of the proteases MMP20 and—subsequently—kallikrein 4 (KLK4) (Extended Data Fig. 4, Methods). The peptide populations are also enriched in N and C termini that correspond to known MMP20 and KLK4 cleavage sites (Extended Data Fig. 7, Supplementary Fig. 9). FAM20C phosphorylation and MMP20 and KLK4 proteolysis are the two main processes that occur in vivo during enamel biomineralization. Our observation of products deriving from both processes opens up the possibility of studying in vivo processes of hominin tooth formation across the Pleistocene epoch.

Homo antecessor is known only from the Gran Dolina TD6 assemblage in Atapuerca[9]. Its relationship with other European Middle Pleistocene fossils is heavily debated[3–5,26,27]. It remains contentious as to whether *H. antecessor* represents the last common ancestor of *H. sapiens*, Neanderthals and Denisovans[9], or whether it represents a sister lineage to the last common ancestor of these species[28,29]. We address this issue by conducting phylogenetic analyses on the basis of our ancient protein sequences from *H. antecessor* (ATD6-92), a panel of present-day great ape genomes and protein sequences translated from archaic hominin genomes (Methods).

We built several phylogenetic trees using maximum likelihood and Bayesian methods (Fig. 2a, Supplementary Figs. 13–16). In these trees, the *H. antecessor* sequence represents a sister taxon that is closely related to, but not part of, the group composed of Late Pleistocene hominins for which molecular data are available (Fig. 2a, Supplementary Figs. 13, 15, 16). The enamel protein sequences do not resolve the relationships between *H. sapiens*, Neanderthals and Denisovans owing to the low number of informative single amino acid polymorphisms. However, pairwise divergence of the amino acid sequences between *H. antecessor* and the clade containing *H. sapiens*, Neanderthals and the Denisovan is larger than the divergence between the members of this clade (Fig. 3b, Supplementary Fig. 12, Supplementary Information). The concatenated gene tree may be subjected to incomplete lineage sorting, and we have too little sequence data to discard this possibility at the moment. However, if we use the concatenation of available gene trees as a best guess for the population tree, and assume that such a population tree is a good descriptor of the relationships among ancient hominins, then our results support the placement of *H. antecessor* as a closely related sister taxon of the last common ancestor of *H. sapiens*, Neanderthals and Denisovans. The phylogenetic position of *H. antecessor* agrees with a divergence of the *H. sapiens* and Neanderthal + Denisovan lineages between 550 and 765 ka[30,31], as ATD6-92 has been dated to 772–949 ka[11]. This is further supported by recent reconsiderations of the morphology of *H. antecessor* in relation to Middle and Late Pleistocene hominins[29].

Homo antecessor has tentatively been proposed as the last common ancestor of Neanderthals and modern humans[9]. The similarities observed between the modern-like mid-facial topography of *H. antecessor* and *H. sapiens*—including a modern pattern of coronal orientation of the infraorbital surface, the sloping and directionality of this plane, as well as the anterior flexion of the maxillary surface and arching of the zygomatic-alveolar crest—were key in this proposal[9,32]. Additional studies of the face of ATD6-69 have confirmed that *H. antecessor* exhibits the oldest known modern-like face in the fossil record[12,13]. The phylogenetic placement of *H. antecessor* implies that this modern-like face—as represented by *H. antecessor*—must have a considerably deep ancestry in the genus *Homo*. Findings made between 2003 and 2005 have shown that the *H. antecessor* hypodigm includes some features that were previously considered Neanderthal autapomorphies[28]. Our results suggest that these features appeared in Early Pleistocene hominins, and were retained by Neanderthals and lost by modern humans.

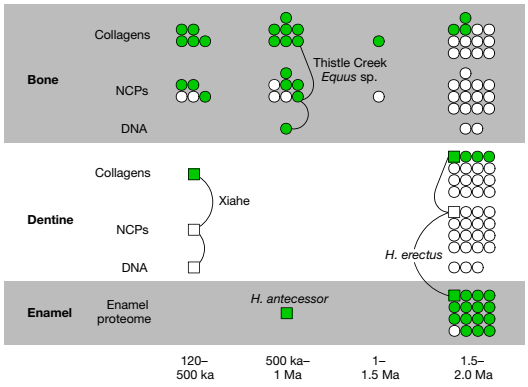


Fig. 3 | Skeletal proteome preservation in the Middle and Early Pleistocene epoch (0.12–2.6 Ma). For each sample, the presence (green) or absence (blank) of endogenous DNA, collagens, non-collagenous proteins (NCPs) or an enamel proteome is given. Only samples for which mammalian proteomes are published are considered[6–8,35–38]. Hominin samples are indicated with squares, other mammalian samples are indicated with circles. Selected specimens have their separate molecular components joined, and are named. Xiahe refers to the Xiahe mandible[7]; the Thistle Creek *Equus* refers to a horse metapodial from the Canadian permafrost[38].

By contrast, the phylogenetic tree built with the *H. erectus* specimen from Dmanisi has only moderate resolution (Extended Data Fig. 8, Supplementary Fig. 11), despite deeper shotgun protein sequencing for this specimen (Extended Data Table 1). This partly inconclusive result might be due to the shorter average peptide lengths compared to the Atapuerca *H. antecessor* specimen (Extended Data Fig. 6d, Methods) and an absence of uniquely segregating single amino acid polymorphisms (Supplementary Table 9). Although our *H. erectus* data from Dmanisi demonstrate that ancient hominin proteins can be reliably obtained from the Early Pleistocene epoch, they also highlight the current limits of ancient protein analysis when applied to the phylogenetic placement of Early Pleistocene hominin remains.

Our dataset provides a unique molecular resource of hominin biomolecular sequences from Early and Middle Pleistocene hominins, and represents—to our knowledge—the oldest ancient hominin proteomes presented to date. Comparison of hominin and fauna proteomes from different skeletal tissues reveals that the dental enamel proteome outlasts dentine and bone proteome preservation (Fig. 3). Here the prolonged survival of hominin enamel proteomes is exploited to show that *H. antecessor* represents a hominin taxon closely related to the last common ancestor of *H. sapiens*, Neanderthals and Denisovans. In addition, our datasets demonstrate that in vivo proteome modifications, such as serine phosphorylation, survive over time scales of hundreds of thousands of years. Current research therefore suggests that dental enamel, the hardest tissue in the mammalian skeleton, is the material of choice for the analysis of hominin evolution in deep time.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2153-8.

1. Gabunia, L. et al. Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019–1025 (2000).
2. Zhu, Z. et al. Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago. *Nature* **559**, 608–612 (2018).
3. Stringer, C. The origin and evolution of *Homo sapiens*. *Phil. Trans. R. Soc. Lond. B* **371**, 20150237 (2016).
4. Hublin, J. J. The origin of Neandertals. *Proc. Natl Acad. Sci. USA* **106**, 16022–16027 (2009).
5. Rightmire, G. Human evolution in the Middle Pleistocene: the role of *Homo heidelbergensis*. *Evol. Anthropol.* **6**, 218–227 (1998).
6. Cappellini, E. et al. Early Pleistocene enamel proteome from Dmanisi resolves *Stephanorhinus* phylogeny. *Nature* **574**, 103–107 (2019).
7. Chen, F. et al. A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409–412 (2019).
8. Welker, F. et al. Enamel proteome shows that *Gigantopithecus* was an early diverging pongine. *Nature* **576**, 262–265 (2019).
9. Bermúdez de Castro, J. M. et al. A hominid from the lower Pleistocene of Atapuerca, Spain: possible ancestor to Neandertals and modern humans. *Science* **276**, 1392–1395 (1997).
10. Carbonell, E. et al. Lower Pleistocene hominids and artifacts from Atapuerca-TD6 (Spain). *Science* **269**, 826–830 (1995).
11. Duval, M. et al. The first direct ESR dating of a hominin tooth from Atapuerca Gran Dolina TD-6 (Spain) supports the antiquity of *Homo antecessor*. *Quat. Geochronol.* **47**, 120–137 (2018).
12. Freidline, S. E., Gunz, P., Harvati, K. & Hublin, J.-J. Evaluating developmental shape changes in *Homo antecessor* subadult facial morphology. *J. Hum. Evol.* **65**, 404–423 (2013).
13. Lacruz, R. S. et al. Facial morphogenesis of the earliest Europeans. *PLoS One* **8**, e65199 (2013).
14. Ferring, R. et al. Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85–1.78 Ma. *Proc. Natl Acad. Sci. USA* **108**, 10432–10436 (2011).
15. Lordkipanidze, D. et al. A complete skull from Dmanisi, Georgia, and the evolutionary biology of early *Homo*. *Science* **342**, 326–331 (2013).
16. Stewart, N. A., Gerlach, R. F., Gowland, R. L., Gron, K. J. & Montgomery, J. Sex determination of human remains from peptides in tooth enamel. *Proc. Natl Acad. Sci. USA* **114**, 13649–13654 (2017).
17. Tiwary, S. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
18. Castiblanco, G. A. et al. Identification of proteins from human permanent erupted enamel. *Eur. J. Oral Sci.* **123**, 390–395 (2015).
19. Asaka, T. et al. Type XVII collagen is a key player in tooth enamel formation. *Am. J. Pathol.* **174**, 91–100 (2009).
20. Porto, I. M., Laure, H. J., de Sousa, F. B., Rosa, J. C. & Gerlach, R. F. New techniques for the recovery of small amounts of mature enamel proteins. *J. Archaeol. Sci.* **38**, 3596–3604 (2011).
21. Gasse, B., Chiari, Y., Silvent, J., Davit-Béal, T. & Sire, J.-Y. Amelotin: an enamel matrix protein that experienced distinct evolutionary histories in amphibians, sauropsids and mammals. *BMC Evol. Biol.* **15**, 47 (2015).
22. Demarchi, B. et al. Protein sequences bound to mineral surfaces persist into deep time. *eLife* **5**, e17092 (2016).
23. Tagliabracci, V. S. et al. Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **336**, 1150–1153 (2012).
24. Hu, J. C. C., Yamakoshi, Y., Yamakoshi, F., Krebsbach, P. H. & Simmer, J. P. Proteomics and genetics of dental enamel. *Cells Tissues Organs* **181**, 219–231 (2005).
25. Glimcher, M. J., Cohen-Solal, L., Kossiva, D. & de Ricqles, A. Biochemical analyses of fossil enamel and dentin. *Paleobiology* **16**, 219–232 (1990).
26. Wagner, G. A. et al. Radiometric dating of the type-site for *Homo heidelbergensis* at Mauer, Germany. *Proc. Natl Acad. Sci. USA* **107**, 19726–19730 (2010).
27. Martinón-Torres, M. et al. Dental evidence on the hominin dispersals during the Pleistocene. *Proc. Natl Acad. Sci. USA* **104**, 13279–13282 (2007).
28. Bermúdez de Castro, J. M., Martinón-Torres, M., Arsuaga, J. L. & Carbonell, E. Twentieth anniversary of *Homo antecessor* (1997–2017): a review. *Evol. Anthropol.* **26**, 157–171 (2017).
29. Gómez-Robles, A., Bermúdez de Castro, J. M., Arsuaga, J.-L., Carbonell, E. & Polly, P. D. No known hominin species matches the expected dental morphology of the last common ancestor of Neanderthals and modern humans. *Proc. Natl Acad. Sci. USA* **110**, 18196–18201 (2013).
30. Meyer, M. et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016).
31. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
32. Lacruz, R. S. et al. The evolutionary history of the human face. *Nat. Ecol. Evol.* **3**, 726–736 (2019).
33. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
34. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
35. Welker, F. et al. Middle Pleistocene protein sequences from the rhinoceros genus *Stephanorhinus* and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ* **5**, e3033 (2017).
36. Hill, R. C. et al. Preserved proteins from extinct *Bison latifrons* identified by tandem mass spectrometry; hydroxylysine glycosides are a common feature of ancient collagen. *Mol. Cell. Proteomics* **14**, 1946–1958 (2015).
37. Wadsworth, C. & Buckley, M. Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. *Rapid Commun. Mass Spectrom.* **28**, 605–615 (2014).
38. Orlando, L. et al. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Site location and specimen selection

Recent human control specimens. We analysed Ø1952, a human premolar recovered in an archaeological excavation in Copenhagen (Almindeligt Hospital Kirkegård, excavated in 1952, from kisse '2'). The tooth is approximately three centuries old, as the cemetery was in use from approximately AD 1600 to approximately AD 1800, and originates from a male individual. We also re-analysed previously published data[16] related to specimens that are dated to between approximately 5,700 and 200 years ago; of these specimens, we took SK339 as a recent example in our comparative figures (a male individual from Fewston (UK) dated to the nineteenth century AD).

Atapuerca. One fragmentary permanent lower left first or second molar (ATD6-92; field number and museum accession number at CENIEH) was used for ancient protein analysis (Extended Data Fig. 2a, Supplementary Information). ATD6-92 originates from sublevel TD6.2 of the Gran Dolina cave site. Sublevel TD6.2 contains a large number of faunal remains, about 170 hominin fossils and about 830 archaeological artefacts. All hominin specimens from sublevel TD6.2, including ATD6-92, are attributed to *H. antecessor*[9]. ATD6-92 has recently been directly dated through electron spin resonance, laser-ablation inductively coupled plasma mass spectrometry U-series and bulk U-series dating[11]. Together with previous chronological research at the site, these analyses constrain the age of ATD6-92 to 772–949 thousand years old[11].

Dmanisi. One fragmentary permanent upper first molar (D4163; field number and museum accession number at the Georgian National Museum) was used for ancient protein analysis (Extended Data Fig. 2b, Supplementary Information). D4163 derives from layer B1 in excavation block M6 (Dmanisi). Layer B1 at Dmanisi contains one of the richest palaeontological assemblages attributed to the Eurasian Early Pleistocene epoch, including several hominin crania. Below, we refer to these specimens as *H. erectus* (Dmanisi). They represent the earliest hominin fossils outside Africa, and are dated to 1.77 Ma[14]. Faunal material from the site previously demonstrated ancient protein survival for most specimens, but a total absence of ancient DNA[6] (Fig. 3).

Amino acid racemization

Chiral amino acid analysis was undertaken on one Pleistocene sample from the hominin tooth (D4163) to test the endogeneity of the enamel protein through its degradation patterns. The tooth chip was separated into the enamel and dentine portions, and each was powdered with an agate pestle and mortar. All samples were prepared using previously published procedures[39], modified to be optimized for enamel, using a bleach time of 72 h to isolate the intracrystalline protein, demineralization in HCl, KOH neutralization and formation of a biphasic solution through centrifugation[40]. Two subsamples were analysed from each portion: one fraction was directly demineralized and the free amino acids analysed, and the second was treated to release the peptide-bound amino acids, thus yielding the total hydrolysable amino acid fraction. Samples were analysed in duplicate by reversed-phase high-performance liquid chromatography, with standards and blanks analysed alongside samples. During preparative hydrolysis, both asparagine (Asn) and glutamine (Gln) undergo rapid irreversible deamidation to aspartic acid (Asp) and glutamic acid (Glu), respectively[41]. It is therefore not possible to distinguish between the acidic amino acids and their derivatives, and they are reported together as Asx and Glx, respectively. Additional descriptions of the methods, as well as additional results, are given in the Supplementary Information.

Proteomic extraction and nanoLC–MS/MS

Protein extraction. Protein extraction was conducted on enamel samples (from the Atapuerca *H. antecessor*, Dmanisi *H. erectus* and Ø1952) and a dentine sample (Dmanisi), using one of three protocols. In brief, the first extraction method used HCl for demineralization, but included no subsequent reduction, alkylation or digestion. The second extraction method used a more standard approach, in which the pellet left from the demineralization in extraction one was reduced, alkylated and digested with LysC and trypsin. The third extraction method used TFA for demineralization, and had no subsequent reduction, alkylation or digestion. The first and third extraction approaches provided more extensive peptide recovery in ancient enamel proteomes[6] compared to the second extraction approach[42]. Further details can be found in the Supplementary Information and a previous publication[6]. Ø1952 was processed using extraction methods one and three. No proteinase and phosphatase inhibitors were used during extraction, as we assumed that catalytically active enzymes were not present in our specimens and the high acidic conditions during our extraction would have irreversibly denatured any proteases possibly present as contaminants in our reagents. Extended Data Table 1 provides a breakdown of the use of specific extraction methods, hominin samples and hominin tissues.

NanoLC–MS/MS analysis. Shotgun proteomic data were obtained on peptide extracts of both hominins at separate facilities at the Novo Nordisk Centre for Protein Research (University of Copenhagen) and the Proteomics Unit (Centre for Genomic Regulation, Barcelona Institute of Science and Technology). Full peptide elutions were injected, in some cases across replicate runs in both Copenhagen and Barcelona. In brief, samples processed in Copenhagen were suspended in 0.1% trifluoroacetic acid, 5% acetonitrile, and analysed on a Q-Exactive HF or HF-X mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1200 (Thermo Fisher Scientific). The HF or HF-X mass spectrometer was operated in positive ion mode with a nanospray voltage of 2 kV and a source temperature of 275 °C. Data-dependent acquisition mode was used for all mass spectrometric measurements. Full mass spectrometry scans were done at a resolution of 120,000 with a mass range of m/z 300–1,750 and 350–1,400 for the HF and HF-X mass spectrometers, respectively, with detection in the Orbitrap mass analyser. Fragment ion spectra were produced at a resolution of 60,000 via high-energy collision dissociation (HCD) at a normalized collision energy of 28% and acquired in the Orbitrap mass analyser. In addition, test runs for the Dmanisi sample were performed at a shorter gradient (Supplementary Information). In Barcelona, samples were dissolved in 0.1% formic acid and analysed on a LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1000. The mass spectrometer was operated similarly to the parameters stated for the HF and HF-X mass spectrometers in Copenhagen, except the nanospray voltage was 2.4 kV and full mass spectrometry scans with 1 micro scan were used over a mass range of m/z 350–1,500. Further details of the LC–MS/MS analysis can be found in the Supplementary Information.

Proteomic data analysis

Protein sequence database construction. We constructed an initial Hominidae sequence database containing protein sequences of all major and minor enamel proteins derived from all extant great apes, a hylobatid (*Nomascus leucogenys*) and a macaque (*Macaca mulatta*). Additionally, we added protein sequences translated from extinct Late Pleistocene hominins[30,43], and sequences from *Gorilla beringei*, *Pongo pygmaeus* and *Pongo tapanuliensis*[44–46]. For each protein, we reconstructed the protein sequence of ancestral nodes in the Hominidae family through PhyloBot[47] to minimize cross-species proteomic effects[48], and added missing isoform variation on the basis of the isoforms present for each protein in the human proteome as given by UniProt (Supplementary Information). Furthermore, we downloaded

the entire human reference proteome from UniProt (4 September 2018) for a single separate search to allow matches to proteins previously not encountered in enamel proteomes. To each constructed database, we added a set of known or possible laboratory contaminants to allow for the identification of possible protein contaminants[49].

Proteomic software, settings and false-discovery rate. Raw mass spectrometry data were searched for each specimen and tissue separately in either PEAKS[50] (v.7.5) or MaxQuant[51] (v.1.5.3.30). No fixed modifications were specified in any search. For PEAKS, variable post-translational modifications were set to include proline hydroxylation, glutamine and asparagine deamidation, oxidation (M), phosphorylation (STY), carbamidomethylation (C) and pyroglutamic acid (from Q and E). For MaxQuant, the following variable post-translation modifications were additionally included: ornithine formation (R), oxidation (W), dioxidation (MW), histidine to aspartic acid (H>D), and histidine to hydroxyglutamate. Searches were conducted with unspecific digestion. For PEAKS, precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 0.05 Da, and the false-discovery rate of peptide spectrum matches was set to equal ≤1.0%. For MaxQuant, default settings of 20 ppm for the first search and 4.5 ppm for the final search were used, a fragment mass tolerance of 20 ppm, and peptide spectrum match (PSM) and protein false-discovery rate was set to 1.0%, with a minimum required Andromeda score of 40 for all peptides. Protein matches were accepted with a minimum of two unique peptide matches in either the PEAKS or MaxQuant search. Proteins that conform to these criteria are detailed in Extended Data Table 2. Example MS/MS spectra from the MaxQuant search and overlapping sites of phylogenetic interest (single amino acid polymorphisms) are included as Supplementary Data 1.

Data search iterations. For both the proteomes of Dmanisi and Atapuerca specimens, we conducted two separate initial searches. First, we conducted a search in PEAKS against the entire human proteome. Only standard enamel proteins were identified in these searches, allowing us to continue with more specific searches. For the Dmanisi dentine sample, this first search resulted in a small number of peptides matching to collagen type I only. On the basis of the limited amount of sequence data, no further analysis of the Dmanisi dentine data was therefore conducted. Second, for the enamel data, we conducted a search in PEAKS and MaxQuant against the entire enamel proteome database of all extant and extinct Hominidae. This search was used to observe single amino acid polymorphisms outside the known sequence variation in PEAKS and MaxQuant through the de novo, error-tolerant and/or dependent peptide approaches implemented in each of these search engines. These initial searches indicate overall good protein preservation in both samples and the presence of peptide matches to *Pan*- and *Homo*-derived proteins only.

On the basis of these two initial searches, a novel protein sequence database was used that only includes sequences from the genus *Pan*, the genus *Homo*, their predicted ancestral sequences and novel protein sequences observed for both the Dmanisi or Atapuerca samples. Final searches and subsequent data analysis were conducted against this database using the above search and post-translational modification settings. Positions supported by insufficient spectral data were replaced by 'X', in resulting peptide alignments before phylogenetic analysis.

Data analysis of Ø1952 and the previously published[16] dataset was conducted only in MaxQuant against a database restricted to *H. sapiens*. All other search settings and database restrictions were similar between these two recent human controls and the ancient hominin proteomes.

Peptide sequence and single amino acid polymorphism validation. To validate the PSMs covering single amino acid polymorphisms of interest, we performed peptide spectrum intensity prediction and validation on our dataset using wiNNer[17]. Data from the ancient specimens (Dmanisi *H. erectus* and Atapuerca *H. antecessor*) were divided into a subset that contained phylogenetically informative peptide sequences and a larger subset that did not contained these peptides. A training dataset was prepared by taking a subset of the latter peptides, and adding a previously published dataset of enamel proteomes from Dmanisi fauna[6]. We built two models, one for HCD +2 spectra and one for HCD +3 spectra. We took into account the large number of variable modifications observed in our ancient enamel proteomes, and split the retained data for each model into subsets for training, validation and testing (80:10:10). We then obtained Pearson correlation coefficients for the predicted and true fragment intensities in the test dataset and the phylogenetically informative spectra. The architecture of wiNNer was built using Keras (version 2.0.8; https://keras.io) and Tensorflow (version 1.3.0). The wiNNer analysis indicated close correspondence between predicted and true fragment ion intensities (Pearson correlation coefficient medians between 0.85 and 0.76 for different subsets of the data), indicating adequate peptide sequence identification for all our peptides, including phylogenetically informative positions and the variable post-translational modifications. The wiNNer model can be accessed on GitHub (https://github.com/cox-labs/wiNNer.git). Additional methodological details of the wiNNer architecture are given in the Supplementary Information.

Protein damage analysis. Ancient proteins can be modified diagenetically in a variety of ways compared to their modern counterparts. We quantify glutamine and asparagine deamidation following a previously publication[42] for MaxQuant output, based on MS1 spectral intensities and protein-based bootstrapping (1,000 bootstraps). Further details can be found in the previous publication[42]. We observe that both glutamines and asparagines are almost all deamidated to glutamic acid and aspartic acid, respectively (Extended Data Fig. 6a–c). In addition, peptide length distributions were obtained for datasets presented here and elsewhere[6,8], demonstrating a shortening of average peptide length and overall peptide length distributions for older samples (Extended Data Fig. 6d).

Protein in vivo modification analysis. The existing literature on enamel and enamel proteome biomineralization describes three processes that are key to the maturation of the enamel proteome: protein hydrolysis by MMP20 and KLK4[52–55], in vivo phosphorylation of serine residues[6,8,23] and expression of different isoforms of AMELX, AMBN and AMTN[52,55,56]. We sought to explore the presence of both in vivo protein hydrolysis and serine phosphorylation modifications in our Pleistocene hominin proteomes.

For protein hydrolysis by MMP20 and KLK4, we made use of the Atapuerca digestion-free dataset and the described locations of AMBN, AMELX and AMELY, and ENAM cleavage by MMP20 and KLK4[52–55]. We compared the experimentally observed cleavage sites to a random cleavage model of each protein separately and tested whether the cleavage sites are present in a larger portion of PSMs in the ancient sample. Here we can indeed show an increased presence of PSMs with termini at, or close to, known MMP20 and KLK4 cleavage locations (Extended Data Fig. 7). This corresponds with our observation that protein regions with continuous sequence coverage correspond to known proteolytic fragments after MMP20 and KLK4 activity (Extended Data Fig. 4).

Phosphorylation of serines (S), threonines (T) and tyrosines (Y) was assessed using Icelogo[57] sequence motif analysis. This analysis was based on the MaxQuant results, from which only identified phosphorylation sites with a localization probability of ≥0.95 were selected. STY sites with no phosphorylation or localization probabilities ≤ 0.95 were taken as the non-phosphorylated background, and a sequence motif window of 7 amino acids on either side of the STY was selected. Sequence motif analysis indicates a strong preference for the phosphorylation of S with a glutamic acid (E) on the +2 position (S-X-E motif) (Fig. 1a, b) in both hominin enamel proteomes. This substrate motif

and the S-X-phosphorylated S motif are recognized by the kinase FAM20C, which is known to be active in vivo on extracellular proteins involved in biomineralization[23], and has previously been reported for ancient, non-hominin enamel proteomes as well[6,8].

To compare phosphorylation occupancy between the Dmanisi and Atapuerca enamel proteomes, we performed a separate MaxQuant database search (Supplementary Information) and restricted our analyses to amino acid positions covered by phosphorylated and non-phosphorylated peptides, observed in both hominins and quantified through label-free quantification.

Phylogenetic analysis

Comparison between the ancient protein sequences and modern reference proteins. We compared the reconstructed ancient protein sequences from the Dmanisi *H. erectus* and Atapuerca *H. antecessor* with protein sequences from great apes[44,46], three Neanderthals[31,43,58], a Denisovan[59] and a panel of present-day humans, including 256 samples from the Simons Genome Diversity Panel[33] and 41 high-coverage individuals from the 1000 Genomes Project[34]. Altogether, our reference data represent worldwide human and great ape variation (Supplementary Tables 7, 8). Additionally, we included protein sequences from macaque (*M. mulatta*) and gibbon (*N. leucogenys*) to root phylogenetic trees. The protein sequences were retrieved from the UniProt database or reconstructed from the reference whole-genome sequences as described in Supplementary Methods.

The ancient and reference protein sequences were aligned using mafft[60]. We aligned the sequences of each protein separately and obtained an alignment for each of the ancient individuals independently (Supplementary Table 9). The isobaric amino acids leucine (L) and isoleucine (I) cannot be distinguished with the experimental procedure used for this study. Therefore, we have to take the following precautions to avoid unintentional sequence differences. If either I or L was present at a specific amino acid position in the reference protein sequences, we replaced all corresponding amino acids in the ancient protein sequences with the amino acid that is present. Alternatively, if both amino acids are present in the reference protein sequence, we replace all I to L for all sequences. We used sequence information for seven proteins (ALB, AMBN, AMELX, AMELY, COL17α1, ENAM and MMP20) for the *H. antecessor* individual and six proteins for the *H. erectus* individual (ALB, AMBN, AMELX, COL17α1, ENAM and MMP20) with a total of 22.08% and 22.14% non-missing sites, respectively (Supplementary Table 9). We were able to recover a unique single amino acid polymorphism for *H. antecessor*; however, for *H. erectus* no unique single amino acid polymorphism was detected (Supplementary Tables 9–11, Supplementary Figs. 10–12).

Phylogenetic reconstruction. We built phylogenetic trees using our protein sequence alignments following three approaches: a maximum likelihood approach using PhyML v.3[61], and two Bayesian approaches using mrBayes[62] and BEAST[63].

For the maximum likelihood approach, we built maximum likelihood trees for each protein independently and for a concatenated alignment consisting of all of the available protein sequences for each of the ancient samples (Supplementary Figs. 13, 14). We used PhyML v.3 and the parameters described in the Supplementary Information section 2.3.5a to build and optimize the tree topologies, branch length and substitutions rates for each of the alignments. Support for each bipartition was obtained based on 100 non-parametric bootstrap replicates. We evaluated the effect of significant missingness in the ancient samples on the inferred topology. Finally, we looked at the effect of varying which of the subset of present-day human samples was included in the tree (Supplementary Information section 2.3.5b, c).

For the Bayesian approach using mrBayes, to assess the robustness of the maximum likelihood inference results, we performed Bayesian phylogenetic inference on the basis of the concatenated alignments using mrBayes 3.2 and the parameters described in Supplementary Information section 2.3.5d (Extended Data Fig. 8, Supplementary Fig. 16). Bayesian inference was performed using the CIPRES Science Gateway[64].

For the Bayesian approach using BEAST, we used BEAST 2.5 to obtain a time calibrated tree for the seven proteins used for *H. antecessor*. For this analysis, we used concatenated alignments including the Neanderthals, the Denisovan, seven randomly chosen *H. sapiens* individuals and a single individual per great ape species. The alignment was partitioned by gene and a coalescent constant population model was used for the tree prior. The dates of the ancient samples included in the analysis (Vindija Neanderthal, 52 ka[58]; Altai Neanderthal, 112 ka[31]; Denisovan, 72 ka[59] and *H. antecessor*, 860.5 ka[11]) were used as tip dates for calibration. For each partition, we used the Jones–Taylor–Thornton substitution model with four categories for the gamma parameter, for which we allowed the Markov chain Monte Carlo chain to sample the shape of the gamma distribution (with an exponentially distributed prior) and assigned independent clock models. Additionally, we set a prior for the divergence time of great apes to 23.85 ± 2.5 Ma (normally distributed)[65], and rooted the tree using the macaque (*M. mulatta*). The overall topology of the tree was estimated for the seven partitions jointly. The convergence of the algorithm was assessed using Tracer v.1.7.0[66]. Finally, we repeated this analysis with 100 alignments, each of them consisting of 7 present-day humans chosen randomly. Although the topology within the clade consisting of present-day humans, Neanderthals and Denisovan was not consistent across the replicates, 99 of the replicates consistently place the *H. antecessor* sequence as an outgroup to this clade (Fig. 2a).

Further details on phylogenetic analysis and results can be found in the Supplementary Information. Example MS/MS spectra from the MaxQuant search and overlapping sites of phylogenetic interest (single amino acid polymorphisms) for both hominins are included as Supplementary Data 1.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD014342. Generated ancient protein consensus sequences used for phylogenetic analysis for *H. antecessor* (Atapuerca) and *H. erectus* (Dmanisi) hominins can be found in the Supplementary Data 2, which is formatted as a .fasta file. Full protein sequence alignments used during phylogenetic analysis can be accessed via Figshare (https://doi.org/10.6084/m9.figshare.9927074). Amino acid racemization data are available online through the NOAA database. The wiNNer model can be accessed on GitHub (https://github.com/cox-labs/wiNNer.git).

39. Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quat. Geochronol.* **3**, 2–25 (2008).
40. Dickinson, M., Lister, A. M. & Penkman, K. E. H. A new method for enamel amino acid racemization dating: a closed system approach. *Quat. Geochronol.* **50**, 29–46 (2019).
41. Hill, R. L. Hydrolysis of proteins. *Adv. Protein Chem.* **20**, 37–107 (1965).
42. Mackie, M. et al. Palaeoproteomic profiling of conservation layers on a 14th century italian wall painting. *Angew. Chem. Int. Ed. Engl.* **57**, 7369–7374 (2018).
43. Castellano, S. et al. Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl Acad. Sci. USA* **111**, 6666–6671 (2014).
44. de Manuel, M. et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).
45. Nater, A. et al. Morphometric, behavioral, and genomic evidence for a new orangutan species. *Curr. Biol.* **27**, 3487–3498.e10 (2017).
46. Prado-Martinez, J. et al. Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).

Article

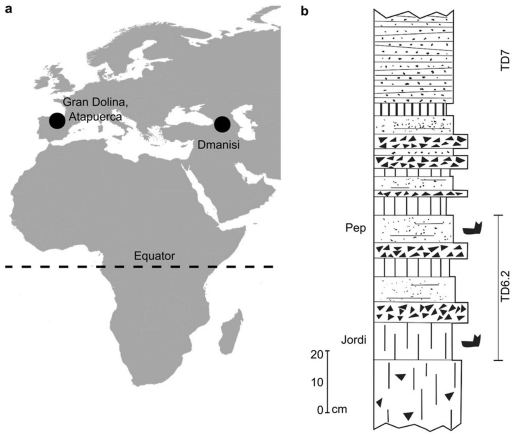
47. Hanson-Smith, V. & Johnson, A. PhyloBot: a web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLOS Comput. Biol.* **12**, e1004976 (2016).
48. Welker, F. Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. *BMC Evol. Biol.* **18**, 23 (2018).
49. Hendy, J. et al. A guide to ancient protein studies. *Nat. Ecol. Evol.* **2**, 791–799 (2018).
50. Zhang, J. et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).
51. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
52. Chun, Y. H. P. et al. Cleavage site specificity of MMP-20 for secretory-stage ameloblastin. *J. Dent. Res.* **89**, 785–790 (2010).
53. Yamakoshi, Y., Hu, J. C. C., Fukae, M., Yamakoshi, F. & Simmer, J. P. How do enamelysin and kallikrein 4 process the 32-kDa enamelin? *Eur. J. Oral Sci.* **114** (Suppl 1), 45–51, 93–95, 379–380 (2006).
54. Iwata, T. et al. Processing of ameloblastin by MMP-20. *J. Dent. Res.* **86**, 153–157 (2007).
55. Nagano, T. et al. Mmp-20 and Klk4 cleavage site preferences for amelogenin sequences. *J. Dent. Res.* **88**, 823–828 (2009).
56. Fukae, M. et al. Primary structure of the porcine 89-kDa enamelin. *Adv. Dent. Res.* **10**, 111–118 (1996).
57. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 (2009).
58. Prüfer, K. et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
59. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, *222–226* (2012).
60. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
61. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
62. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
63. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
64. Miller, M. A., Pfeiffer, W. & Schwartz, T. in *Gateway Computing Environments Workshop (GCE)* 1–8 (New Orleans, 2010).
65. Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T. & Schierup, M. H. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat. Ecol. Evol.* **3**, 286–292 (2019).
66. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

Acknowledgements F.W. is supported by a Marie Skłodowska Curie Individual Fellowship (no. 795569). E. Cappellini was supported by VILLUM FONDEN (no. 17649). E.W. is supported by the Lundbeck Foundation, the Danish National Research Foundation, the Novo Nordisk Foundation, the Carlsberg Foundation, KU2016 and the Wellcome Trust. Without the effort of the members of the Atapuerca research team during fieldwork, this work would have not been possible; we make a special mention of J. Rosell, who supervises the excavation of the TD6 level. The research of the Atapuerca project has been supported by the Dirección General de Investigación of the Ministerio de Ciencia, Innovación y Universidades (grant numbers PGC2018-093925-B-C31, C32, and C33); field seasons are supported by the Consejería de Cultura y Turismo of the Junta de Castilla y León and the Fundación Atapuerca. We acknowledge The Leakey Foundation through the personal support of G. Getty (2013) and D. Crook (2014–2016, 2018, and 2019) to M.M.-T., as well as F.W. (2017). Restoration and conservation work on the material have been carried out by P. Fernández-Colón and E. Lacasa from the Conservation and Restoration Area of CENIEH-ICTS and L. López-Polín from IPHES. The picture of the specimen ATD6-92 was made by M. Modesto-Mata. E. Cappellini, J.C., J.V.O. and P. Gutenbrunner are supported by the Marie Skłodowska-Curie European Training Network (ETN) TEMPERA, a project funded by the European Union's Framework Program for Research and Innovation Horizon 2020 (grant agreement no. 722606). Amino acid analyses were undertaken thanks to the Leverhulme Trust (PLP-2012-116) and NERC (NE/K500987/1). T.M.-B. is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social 'La Caixa' and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). C.L.-F. is supported by a FEDER-MINECO grant (PGC2018-095931-B-100). M.K. was supported by the Postdoctoral Junior Leader Fellowship Programme from 'la Caixa' Banking Foundation (LCF/BQ/PR19/11700002). M.M. is supported by the Danish National Research Foundation award PROTEIOS (DNRF128). Work at the Novo Nordisk Foundation Center for Protein Research is funded in part by a donation from the Novo Nordisk Foundation (grant number NNF14CC0001). The CRG/UPF Proteomics Unit is part of the Spanish Infrastructure for Omics Technologies (ICTS OmicsTech) and it is a member of the ProteoRed PRB3 consortium, which is supported by grant PT17/0019 of the PE I+D+i 2013-2016 from the Instituto de Salud Carlos III (ISCIII) and ERDF. We acknowledge support from the Spanish Ministry of Science, Innovation and Universities, 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208, and 'Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya' (2017SGR595). D.L. and A.M. are supported by the John Templeton Foundation (no. 52935) and by the Shota Rustaveli National Science Foundation of Georgia (no. FR-18-27262). We thank M. L. Schjellerup Jørkov for providing specimen Ø1952.

Author contributions E. Cappellini, E.W., J.M.B.d.C., D.L., C.L.-F. and F.W. designed the study. E. Cappellini, M.M., F.W., J.R.-M., R.R.J.-C., M.R.D., C.C. and M.d.M. performed experiments. E. Cappellini, A.M., J.L.A., E. Carbonell, P. Gelabert, E.S., J.C., J.V.O., T.M.-B. and D.L. provided material, reagents or research infrastructure. F.W., J.R.-M., P. Gutenbrunner, S.T., E. Cappellini, F.R., M.M.-T., J.M.B.d.C., M.K., M.R.D., C.L.-F. and K.P. analysed data. F.W., E. Cappellini and J.M.B.d.C. wrote the manuscript with input from all other authors.

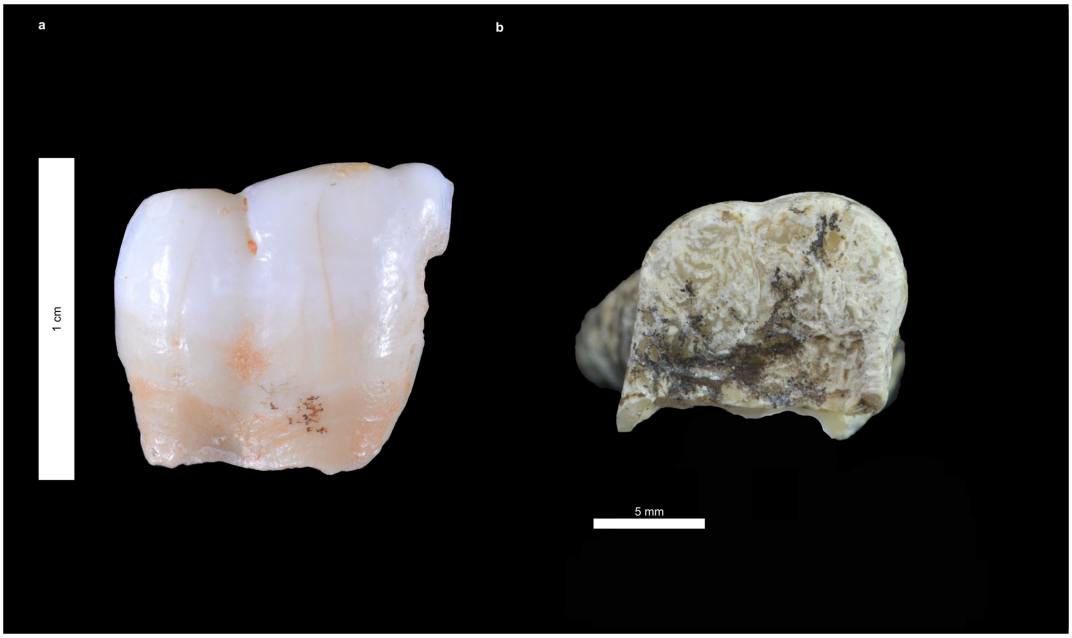
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2153-8.
Correspondence and requests for materials should be addressed to F.W., J.M.B.d.C., E.W. or E.C.
Reprints and permissions information is available at http://www.nature.com/reprints.

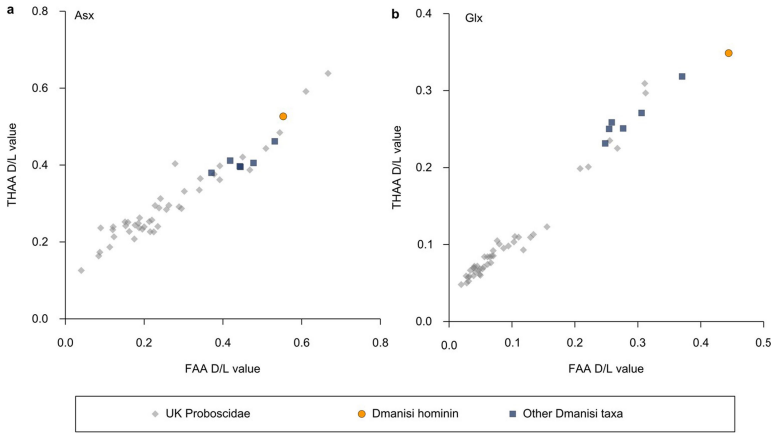


Extended Data Fig. 1 | Location and stratigraphy of the hominin fossils studied. a, Geographic location of Gran Dolina and Dmanisi. Base map was generated using public domain data from www.naturalearthdata.com. **b**, Summarized stratigraphic profile of Gran Dolina, including the location of hominin fossils in layers 'Pep' and 'Jordi' of sublevel TD6.2.

Article

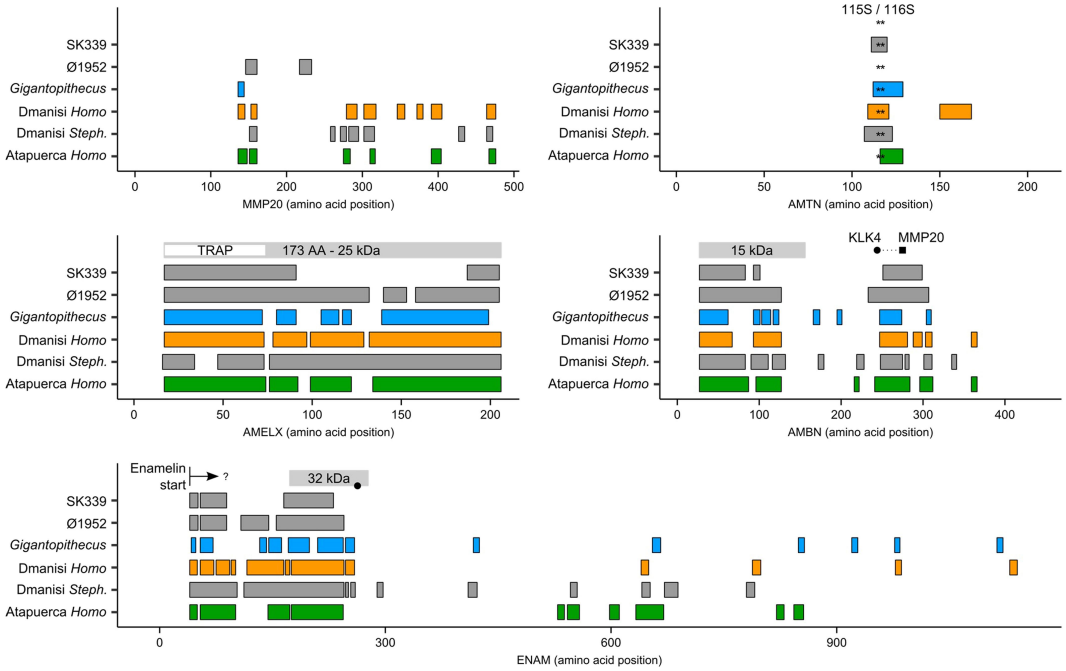


Extended Data Fig. 2 | Hominin specimens studied. a, ATD6-92 in buccal view. The fragment represents a portion of a permanent lower left first or second molar. **b**, D4163 in occlusal view. The specimen is a fragmented right upper first molar. Scale bar differs between **a** and **b**.

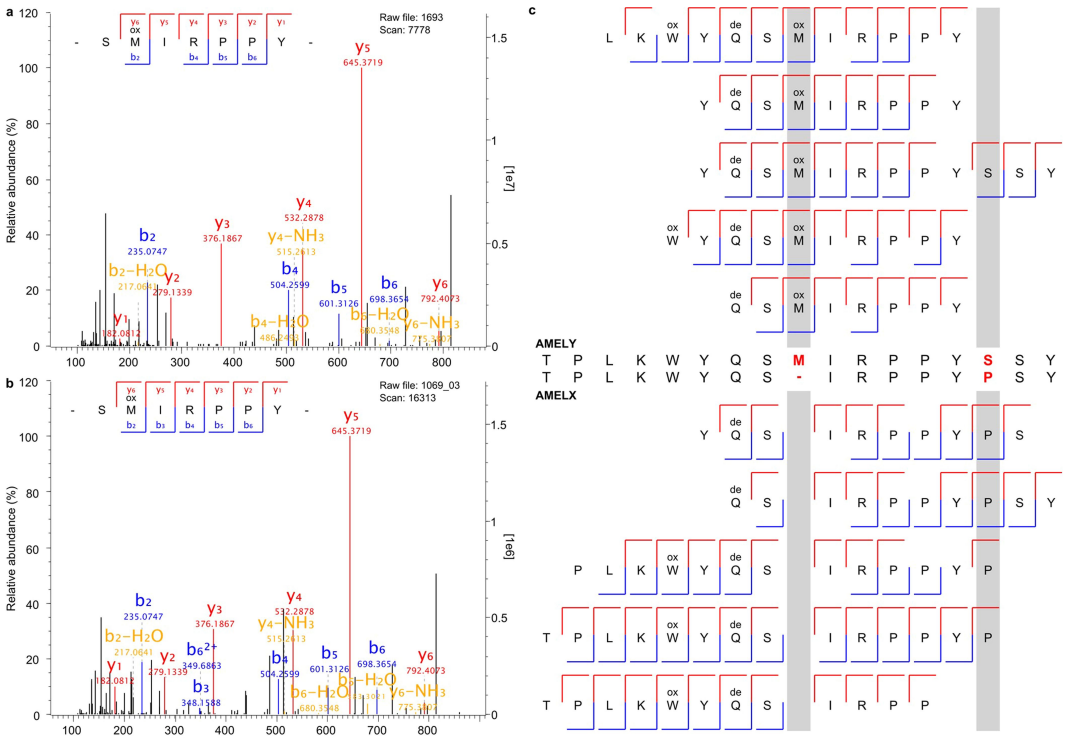


Extended Data Fig. 3 | Amino acid racemization of D4163. a, **b**, The extent of intracrystalline racemization in enamel for the free amino acid (FAA) (*x* axis) fraction and the total hydrolysable amino acids (THAA) (*y* axis) fraction for aspartic acid plus asparagine (here denoted Asx) (**a**), and glutamic acid plus glutamine (here denoted Glx) (**b**), demonstrates endogenous amino acids breaking down within a closed system. The hominin value is displayed in relation to values for enamel samples from other fauna from Dmanisi[6] (blue squares) and a range of previously obtained Pleistocene and Pliocene Proboscidea from the UK[40] (grey diamonds). Fauna are shown for comparison, but different rates in their protein breakdown mean that they will show different extents of racemization. The *x* and *y* axis are on different scales.

Article

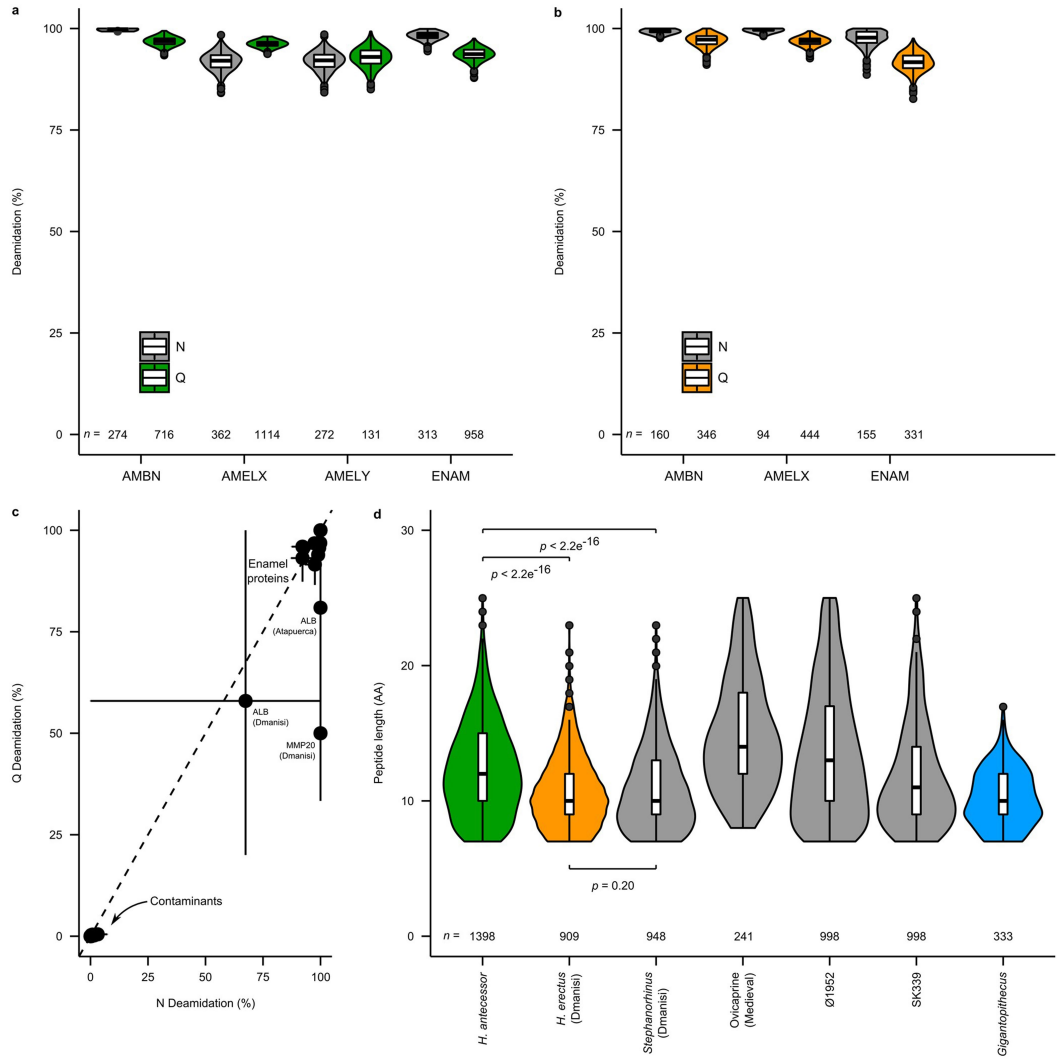


Extended Data Fig. 4 | Sequence coverage for five enamel-specific proteins across Pleistocene samples and recent human controls. For each protein, the bars span protein positions covered, with positions remapped to the human reference proteome. The top row indicates the position of a selection of known MMP20 and KLK4 cleavage products of the enamel-specific proteins AMELX[55], AMBN[52] and ENAM[56]. Several in vivo proteolytic degradation fragments of ENAM share the same N terminus, but have unknown C termini[53]. Dotted line for AMBN indicates a putative cleavage product based on known MMP20 (squares) and KLK4 (circles) in vivo cleavage positions. For AMTN, serines (S) at positions 115 and 116 (indicated by asterisks) are conserved among vertebrates and involved in mineral-binding[21]. Additional cleavage products as well as MMP20 and KLK4 cleavage sites are known in all enamel-specific proteins. SK339[16] and Ø1952 are two recent human control samples (Methods). AA, amino acids; *Steph.*, *Stephanorhinus*[6]; TRAP, tyrosine-rich amelogenin polypeptide.

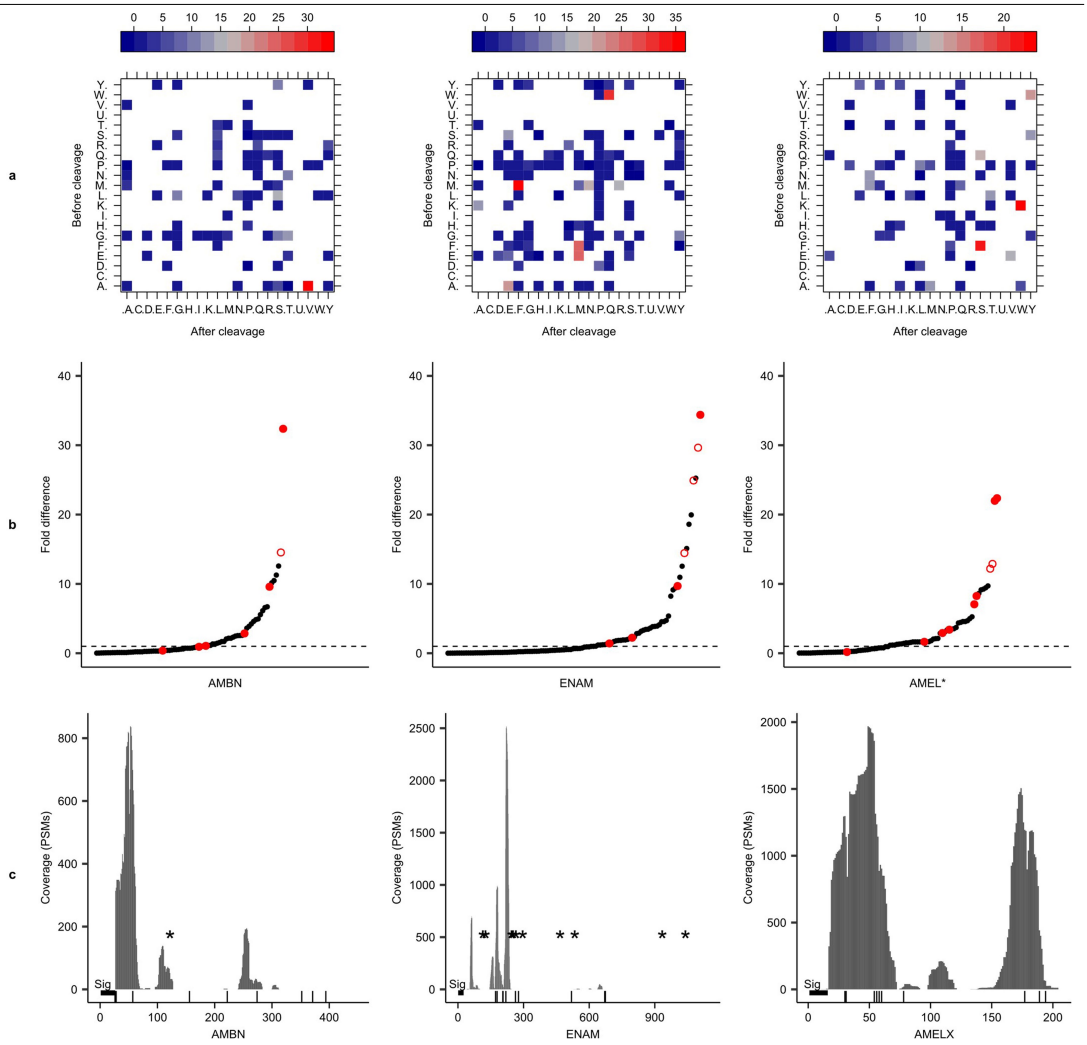


Extended Data Fig. 5 | *Homo antecessor* specimen ATD6-92 represents a male hominin. a, Mass spectrum of an AMELY-specific peptide from the recent human control Ø1952. **b**, Mass spectrum of the same AMELY-specific peptide from *H. antecessor*. **c**, Alignment of a selection of AMELY- and AMELX-specific peptide fragment ion series deriving from *H. antecessor*. The alignment stretches along human AMELX isoform 1, positions 37 to 52 only (Uniprot accession numbers Q99217 (AMELX), Q99218 (AMELY)). See Supplementary Fig. 5 for another example of an AMELY-specific MS2 spectrum.

Article

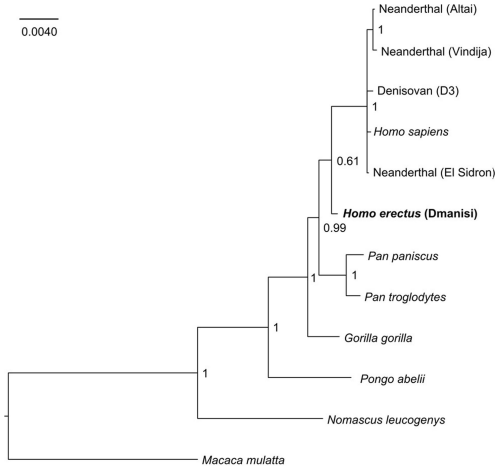


Extended Data Fig. 6 | Enamel proteome damage. a, **b**, Glutamine (Q) and asparagine (N) deamidation of enamel-specific proteins from *H. antecessor* (Atapuerca) (**a**), and *H. erectus* (Dmanisi) (**b**). Values are based on 1,000 bootstrap replications of protein deamidation. **c**, Relationship between mean asparagine (N) and glutamine (Q) deamidation for all proteins in both the Atapuerca and Dmanisi hominin datasets. Error bars represent 95% confidence interval window of 1,000 bootstrap replications of protein deamidation. Dashed line is $x = y$. **d**, Peptide length distribution of *H. antecessor* (Atapuerca), *H. erectus* (Dmanisi), four previously published enamel proteomes[6,8,16] and one additional human Medieval control sample (Ø1952). For **a**, **b** and **d**, the number of peptides (*n*) is given for each violin plot. The box plots within the violin plots define the range of the data (whiskers extend to 1.5× the interquartile range), outliers (black dots, beyond 1.5× the interquartile range), 25th and 75th percentiles (boxes), and medians (centre lines). *P* values of two-sided *t*-tests conducted between sample pairs are indicated. No independent replication of these experiments was performed.



Extended Data Fig. 7 | Survival of in vivo MMP20 and KLK4 cleavage sites in the Atapuerca enamel proteome. a, Experimentally observed cleavage matrices for ameloblastin (AMBN), enamelin (ENAM) and amelogenin (AMELX and AMELY) (Methods). Fold differences are colour-coded by comparing observed PSM cleavage frequencies to a random cleavage matrix for each protein separately[7]. **b**, Fold differences for all observed cleavage pairs per protein. Red filled circles represent MMP20, KLK4 and signal peptide cleavage sites mentioned in the literature[53–56]. Red open circles indicate cleavage sites located up to two amino acid positions away from such sites. **c**, PSM coverage for each protein. The signal peptide (thick horizontal bar labelled 'sig'), known MMP20 and KLK4 cleavage sites (vertical bars), and O- and N-linked glycosylation sites (asterisks) are also indicated. For AMELX, peptide positions for all three known isoforms were remapped to the coordinates of isoform 3, which represents the longest isoform (UniProt accession Q99217-3). The *x* and *y* axes differ between the three panels of **c**.

Article



Extended Data Fig. 8 | Phylogenetic position of D4163 through Bayesian analysis. *Nomascus leucogenys* and *M. mulatta* were used as outgroups.

Extended Data Table 1 | Extraction and mass spectrometry details of analyses conducted on both ancient hominin specimens

Stage Tip number	Tissue	Protein extraction method*	Mass Spectrometer	Mass Spectrometer location	Replicates
***Homo antecessor*, specimen ATD6-92, Atapuerca**					
1069	Enamel	1	QE-HF	Copenhagen	4
1069	Enamel	1	Fusion Lumos	Barcelona	1
***Homo erectus*, specimen D4163, Dmanisi**					
1138	Enamel	1	QE-HF	Copenhagen	2
1141	Enamel	2	QE-HF	Copenhagen	2
1138	Enamel	1	Fusion Lumos	Barcelona	1
1141	Enamel	2	Fusion Lumos	Barcelona	1
1139	Dentine	1	QE-HF	Copenhagen	2
1142	Dentine	2	QE-HF	Copenhagen	2
1139	Dentine	1	Fusion Lumos	Barcelona	1
1142	Dentine	2	Fusion Lumos	Barcelona	1
1386	Enamel	1	QE-HF	Copenhagen	1
1387	Enamel	3	QE-HF	Copenhagen	1
1388	Enamel	1	QE-HF	Copenhagen	1

QE-HF, Q Exactive HF (or HF-X) hybrid quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). Fusion Lumos, LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific).
*Extraction method 1: demineralization in HCl, with no subsequent proteolytic digestion. Extraction method 2: demineralization in HCl, reduction, alkylation and digestion with LysC and trypsin. Extraction method 3: demineralization in TFA, with no subsequent proteolytic digestion. See Supplementary Information for further details.

2.3 Phasing-in quality control in the nucleus

The nuclear proteins are known to aggregate in neurodegenerative disorders such as amyotrophic lateral sclerosis and Huntington's disease. The protein quality control in nucleus is not yet well understood. In this article, the authors studied the protein quality control in the nucleus by using a combination of methods such as fluorescence imaging, proteomics, and biochemical analysis. Proteins enter the nucleus in a folded state, so they do not need chaperons mediated de novo folding. The nuclear proteins are rich in stress-sensitive and metastable proteins, so there should be an effective protein quality control mechanism in the nucleus. The authors showed in the article that metastable nuclear protein, which misfolds upon heat stress, enters the liquid-like granular component (GC) phase in the nucleolus. The GC phase is rich in negatively charged proteins such as nucleophosmin (NPM1) and nucleolin, and it adopts a state of low mobility. Storage of these misfolded proteins in the GC phase effectively prevented the irreversible aggregation, allowing Hsp70-mediated extraction and refolding(or degradation) upon recovery from stress. Disruption of the GC phase causes the formation of stable protein aggregates. Prolonged stress results in a transition of the nucleolar matrix from liquid-like to solid and prevents quality control. To identify the endogenous proteins that enter the GC phase of the nucleolus upon stress, the authors performed GFP-NPM1 pull-down experiments followed by quantitative proteomics. They identified $\sim$200 proteins that are associated with NPM1 specifically upon heat shock (HS) that includes numerous proteins of the nucleoplasm and nucleolus as well as some cytosolic proteins. Thus, the proteins that entered the GC phase constituted a thermally sensitive subproteome. They confirmed that these proteins were enriched in disordered and low complexity regions, hallmarks of metastable structure.

I contributed to the analysis in finding out the enrichment of disorder regions and low complexity regions in the GFP-NPMI associated proteins($\sim$200) upon HS when compared with human proteome. I obtained the human proteome from the Uniprot database[3]. DISOPRED[6], a machine learning-based tool (using SVM) was used to predict disorder regions in the protein. Disorder regions are the regions that lack a fixed or ordered three-dimensional structure. They could be partially or fully unstructured and could form random coils, molten globules, and large multidomain proteins connected by flexible linkers. SEG program[108] was used to mask the low complexity regions with x in the protein sequence. I calculated the distribution of the longest disordered amino

acid sequences and low complexity sequences, as well as the frequency of residues in disordered regions and in low complexity regions, respectively. The result shows that the proteins were enriched in disorder sequences and low complexity sequences, representing the metastable structure when compared to the human proteome. Thus, authors suggest that the nucleolus has chaperone-like properties and can promote nuclear protein maintenance under HS.

F. Frottin, F. Schueder, **Tiwary, S.**, R. Gupta, R. Körner, T. Schlichthaerle, J. Cox, R. Jungmann, F. U. Hartl, and M. S. Hipp. The nucleolus functions as a phase-separated protein quality control compartment. *Science*, 365(6451):342–347, 2019. ISSN 0036-8075. doi: 10.1126/science.aaw9157. URL https://science.sciencemag.org/content/365/6451/342

RESEARCH ARTICLE SUMMARY

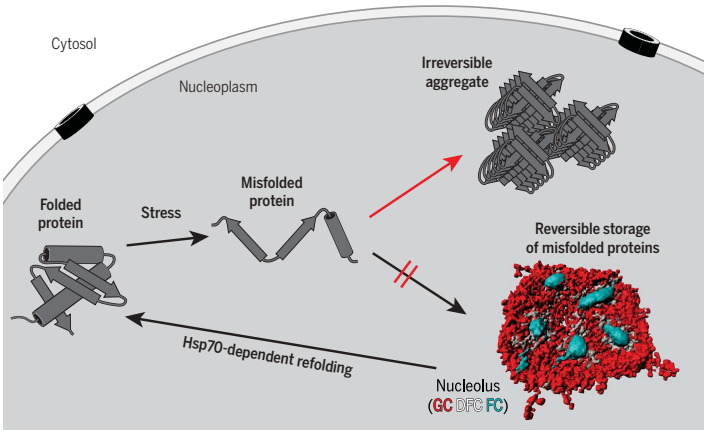
QUALITY CONTROL

The nucleolus functions as a phase-separated protein quality control compartment

F. Frottin, F. Schueder, S. Tiwary, R. Gupta, R. Körner, T. Schlichthaerle, J. Cox, R. Jungmann*, F. U. Hartl*, M. S. Hipp*

INTRODUCTION: Cells have evolved quality control mechanisms that operate under normal growth conditions and during stress to maintain protein homeostasis (proteostasis) and prevent the formation of potentially toxic aggregates. Research in recent decades has identified complex quality control systems in the cytoplasm that mediate protein folding, prevent misfolding, and cooperate in protein degradation with the proteasome and autophagy pathways. Compartment-specific proteostasis networks and stress response pathways have also been described for the endoplasmic reticulum and mitochondria. In contrast, relatively little is known about protein quality control in the nucleus.

Proteins enter the nucleus in a folded state, so chaperone machinery specific for de novo folding is not required. However, the nuclear proteome is rich in stress-sensitive, metastable proteins, which suggests that effective protein quality control mechanisms are in place to ensure conformational maintenance. The nucleus contains several non–membrane-bound subcompartments. The largest of these is the nucleolus, the site of ribosome biogenesis. During stress, Hsp70 and other molecular chaperones accumulate in the nucleolus, presumably to protect unassembled ribosomal proteins against aggregation. The nucleolus consists of liquid-like phases or domains that have differential surface tension and do not intermix. The outermost of these, the granular component (GC), is rich in negatively charged proteins such as nucleophosmin and nucleolin, which, combined with RNA, can undergo phase separation into liquid droplets in vitro, as shown for nucleophosmin.



Inserting misfolded proteins into the nucleolus prevents irreversible aggregation. Upon cell stress, misfolded proteins enter the GC phase of the nucleolus to be stored in a state competent for Hsp70-dependent refolding during recovery. Potentially toxic, irreversible aggregates form when transfer into the nucleolus is prevented. A 3D-rendered high-resolution image of the nucleolus is shown: GC, granular component (red); DFC, dense fibrillar component (white); FC, fibrillar center (cyan).

RATIONALE: Nuclear protein aggregates have been observed in various neurodegenerative disorders such as amyotrophic lateral sclerosis and Huntington's disease, but protein quality control in the nucleus is not well understood. Here, we used a combination of fluorescence imaging, biochemical analyses, and proteomics to investigate the fate of stress-denatured and aberrant proteins in the nucleus, focusing specifically on the role of the nucleolus and its phase-separated nature in protein quality control.

RESULTS: Upon heat stress, misfolded nuclear proteins entered the liquid-like GC phase of the nucleolus, where they associated with proteins including nucleophosmin and adopted a state of low mobility. As a consequence, a fraction of nucleophosmin and nucleolin also converted to a less dynamic state. Storage in the GC phase effectively prevented the irreversible aggregation of misfolded protein species, allowing their extraction and refolding upon recovery from stress in a Hsp70-dependent manner. We identified ~200 different proteins that reversibly partitioned upon stress into the immobile substate of the GC, entering either from the nucleoplasm or from within the nucleolus. Disruption of the GC phase resulted in the formation of stable aggregates of stress-denatured proteins in the nucleoplasm, which exerted toxic effects by sequestering bystander proteins. Notably, the capacity of the nucleolus to store misfolded proteins proved to be limited. Prolonged stress or the uptake of aberrant proteins associated with neurodegenerative diseases led to a transition of the GC phase from a liquid-like to a solid state, with loss of reversibility and nucleolar dysfunction.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaw9157

CONCLUSION: The liquid-like GC phase of the nucleolus functions as a non–membrane-bound protein quality control compartment. It is characterized by a remarkable chaperone-like capacity to temporarily store misfolded proteins, preventing their irreversible aggregation and maintaining them as competent for Hsp70-assisted refolding. Nucleoplasmic proteins exit the nucleolus upon refolding, and nucleolar proteins resume their functional state. Our findings provide an example of how the properties of a non–membrane-bound, phase-separated compartment can be used in protein quality control, a fundamental biological function. ■

The list of author affiliations is available in the full article online.
***Corresponding author. Email: uhartl@biochem.mpg.de (F.U.H.); hipp@biochem.mpg.de (M.S.H.); jungmann@biochem.mpg.de (R.J.)**

RESEARCH ARTICLE

QUALITY CONTROL

The nucleolus functions as a phase-separated protein quality control compartment

F. Frottin[1], F. Schueder[2,3], S. Tiwary[4], R. Gupta[1]*, R. Körner[1], T. Schlichthaerle[2,3], J. Cox[4], R. Jungmann[2,3]†, F. U. Hartl[1,5]†, M. S. Hipp[1,5]†

The nuclear proteome is rich in stress-sensitive proteins, which suggests that effective protein quality control mechanisms are in place to ensure conformational maintenance. We investigated the role of the nucleolus in this process. In mammalian tissue culture cells under stress conditions, misfolded proteins entered the granular component (GC) phase of the nucleolus. Transient associations with nucleolar proteins such as NPM1 conferred low mobility to misfolded proteins within the liquid-like GC phase, avoiding irreversible aggregation. Refolding and extraction of proteins from the nucleolus during recovery from stress was Hsp70-dependent. The capacity of the nucleolus to store misfolded proteins was limited, and prolonged stress led to a transition of the nucleolar matrix from liquid-like to solid, with loss of reversibility and dysfunction in quality control. Thus, we suggest that the nucleolus has chaperone-like properties and can promote nuclear protein maintenance under stress.

Cells have evolved complex quality control mechanisms that operate under normal growth conditions and during stress to maintain protein homeostasis (proteostasis) and prevent the formation of potentially toxic aggregates (*1*–*4*). Subcellular compartments are equipped with specialized stress response pathways (*5*–*7*) and vary in stress vulnerability (*8*–*10*). The nuclear proteome is enriched in proteins containing intrinsically disordered or low-complexity sequences (*11*, *12*). These metastable proteins do not populate a thermodynamically stable folded state and tend to aggregate upon conformational stress (*13*–*15*). Indeed, various neurodegenerative disorders associated with protein aggregation, such as amyotrophic lateral sclerosis (ALS) and Huntington's disease, are characterized by the presence of intranuclear inclusions (*16*–*20*).

The nucleus contains several non–membrane-bound subcompartments (*21*). The largest of these is the nucleolus, which consists of liquid-like phases that do not intermix, giving rise to distinct zones (Fig. 1A and fig. S1, A and B) (*22*). Embedded in the outer granular component (GC) phase is the fibrillar center (FC) for the transcription of ribosomal RNA (RNA polymerase I subunit RPA40 as marker). The FC is surrounded by the dense fibrillar component (DFC), which contains the ribonucleoprotein fibrillarin (FBL) (Fig. 1A and fig. S1, A and B). The GC phase is rich in negatively charged proteins such as nucleophosmin (NPM1) and nucleolin (*23*). NPM1 contains extensive unstructured regions and undergoes liquid-liquid phase separation in vitro (*24*, *25*). During stress, Hsp70 and other molecular chaperones accumulate in the nucleolus, presumably to protect unassembled ribosomal proteins against aggregation (*26*–*28*). Stress-induced transfer of a nuclear model protein to the nucleolus has also been observed (*29*). Here, we found that during stress, misfolded proteins enter the liquid-like GC phase of the nucleolus, where irreversible coaggregation of different misfolded protein species is prevented, allowing Hsp70-mediated extraction and refolding (or degradation) upon recovery from stress. In contrast, disruption of the GC phase causes the formation of stable protein aggregates. Prolonged stress results in a transition of the nucleolar matrix from liquid-like to solid and prevents quality control.

Transfer of misfolded protein to the nucleolus upon stress

To investigate the fate of a nuclear protein as it denatures during heat stress (HS), we generated human embryonic kidney (HEK) 293T cells stably expressing a fusion protein of the thermolabile firefly luciferase and heat-stable green fluorescent protein (GFP) carrying an N-terminal nuclear localization signal (NLS-LG) (fig. S1C). NLS-LG was diffusely distributed in the nucleus. Upon incubation at 43°C (2 hours), a substantial fraction of NLS-LG entered the nucleoli (Fig. 1B). Superresolution imaging (fig. S1A) (*30*, *31*) showed that nucleolar NLS-LG localized to the NPM1-containing GC phase (Fig. 1C and fig. S1D). Transfer of NLS-LG to the nucleolus was prevented by stabilizing luciferase with the substrate analog 2-phenylbenzothiazole (PBT) (Fig. 1B and fig. S1E). Thus, unfolding was a prerequisite for transfer to the nucleolus. Upon recovery from HS, nucleolar NLS-LG redistributed to the nucleoplasm (Fig. 1B), as shown by inhibiting synthesis of new protein (fig. S1F). More than 60% of NLS-LG was degraded during HS (fig. S1G). Notably, the NLS-LG present after recovery showed a higher specific luminescence activity than during HS, indicative of refolding of misfolded protein (fig. S1G).

Hsp70 transferred to nucleoli upon HS (*27*–*29*), even when NLS-LG was stabilized (fig. S1E). Thus, Hsp70 entered the nucleolus either in a complex with endogenous proteins or in free form. Inhibition of the adenosine triphosphatase activity of Hsp70 by the compound VER-155008 (*32*) prevented both Hsp70 and misfolded NLS-LG from exiting the nucleolus during recovery (fig. S2A). Thus, nucleolar Hsp70 is involved in refolding and repartitioning NLS-LG (and presumably other metastable proteins) to the nucleoplasm. Indeed, misfolded cytosolic carboxypeptidase Y*-mCherry (CC*) (*33*) also accumulated in nucleoli when its degradation was inhibited (fig. S2B). Thus, the nucleolus serves as a storage compartment for a subset of misfolded proteins under proteotoxic stress conditions, preserving them in a state competent for refolding or degradation.

Misfolded proteins in the nucleolus have low mobility

We next analyzed the mobility of NLS-LG in the GC phase of the nucleolus by recording fluorescence recovery after photobleaching (FRAP). To compare the mobility of folded and misfolded proteins within the nucleolus, we fused a nucleolar targeting sequence (*34*) to NLS-LG, generating the protein No-LG (fig. S1C). A large fraction of No-LG constitutively localized to the nucleolus in the absence of stress and in the presence of the luciferase stabilizer PBT (Fig. 1D and fig. S2, C and D), thus behaving as a functional nucleolar protein. No-LG in the nucleolus showed complete FRAP (Fig. 1, D and E, and fig. S3A) and a mobility similar to that of the liquid-like GFP-NPM1 (Fig. 1F and fig. S3B) (*22*). HS resulted in a more complete localization of No-LG to the nucleolus, an increase in the nucleolar concentration of No-LG (by a factor of 1.37 ± 0.13, n = 3), and a shift to a markedly reduced mobility (Fig. 1, D and E, and figs. S2, C and D, and S3A). In contrast, the presence of PBT during HS preserved

[1]Department of Cellular Biochemistry, Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany. [2]Research Group "Molecular Imaging and Bionanotechnology," Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany. [3]Faculty of Physics and Center for Nanoscience, Ludwig Maximilian University, D-80539 Munich, Germany. [4]Research Group "Computational Systems Biochemistry," Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany. [5]Munich Cluster for Systems Neurology (SyNergy), D-80336 Munich, Germany.
*Present address: Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark.
†Corresponding author. Email: uhartl@biochem.mpg.de (F.U.H.); hipp@biochem.mpg.de (M.S.H.); jungmann@biochem.mpg.de (R.J.)

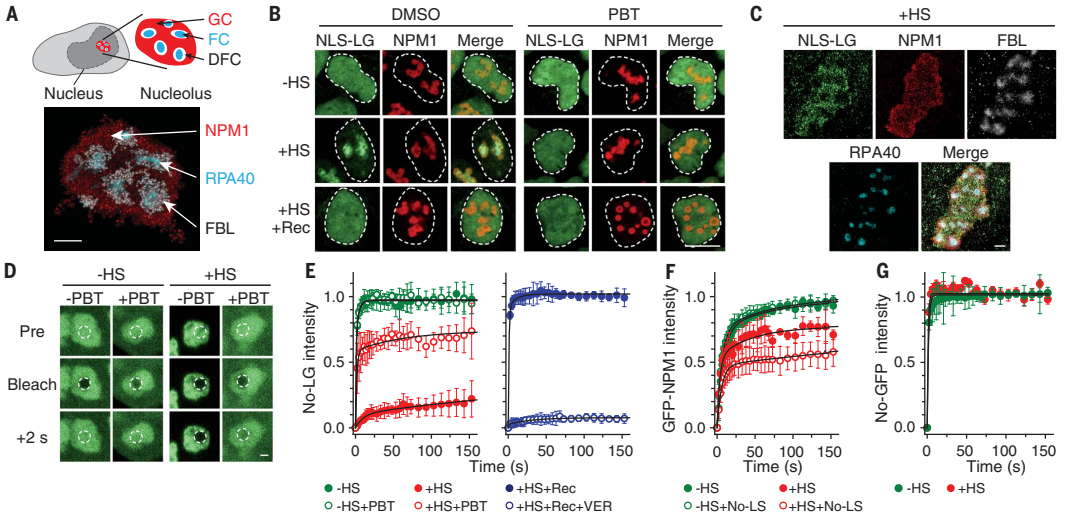


Fig. 1. Misfolded proteins transiently accumulate in the GC phase of the nucleolus during stress. (**A**) Schematic representation and 3D-rendered DNA-PAINT (*30*, *31*) superresolution image of a HeLa cell nucleolus under normal growth conditions. Red, NPM1 (GC); white, FBL (DFC); cyan, RPA40 (FC). See also fig. S1, A and B. (**B**) HEK293T cells stably expressing NLS-LG were treated with dimethyl sulfoxide (DMSO; mock) or PBT before 2 hours of heat stress (+HS), followed by recovery for 2 hours (+HS +Rec). Control cells were maintained at 37°C (−HS). Cells were stained for endogenous NPM1 (red); nuclei are marked by dashed circles. (**C**) Superresolution imaging of HEK293T cells expressing NLS-LG after HS treatment, with staining for GFP, endogenous NPM1, FBL, and RPA40. See fig. S1D for −HS control. (**D**) No-LG in the nucleolus without (−HS) and with (+HS) heat stress in the presence or absence of PBT before bleaching (Pre), immediately after bleaching (Bleach), and 2 s after bleaching. (**E** to **G**) FRAP analysis of No-LG (E), GFP-NPM1 (F), and No-GFP (G). No-LG experiments (E) show PBT treatment (open circles) or DMSO (solid circles) as a control. GFP-NPM1 experiments (F) show cotransfection of No-LS (open circles). For the −HS condition (green), cells were maintained at 37°C during acquisition. For +HS experiments (red), cells were incubated at 43°C for 1 hour before acquisition and maintained at 43°C during acquisition. For the No-LG recovery experiment [(E), right graph, blue], cells were subjected to HS and allowed to recover for 1 hour (+HS +Rec; solid circles), followed by FRAP. Hsp70 was inhibited with VER-155008 before shifting cells to recovery (+HS +Rec +VER; open circles) (*32*). Cycloheximide was present during recovery. The graphs display corrected and normalized FRAP curves with double-exponential fits. Curves represent means ± SD ($n \geq 4$ biological repeats representing at least 12 different cells). The first 150 s after photobleaching are shown. Quantification of No-LG and GFP-NPM1 mobility is shown in fig. S3, A and B, respectively. Scale bars, 1 μm [(A), (C), (D)], 10 μm (B).

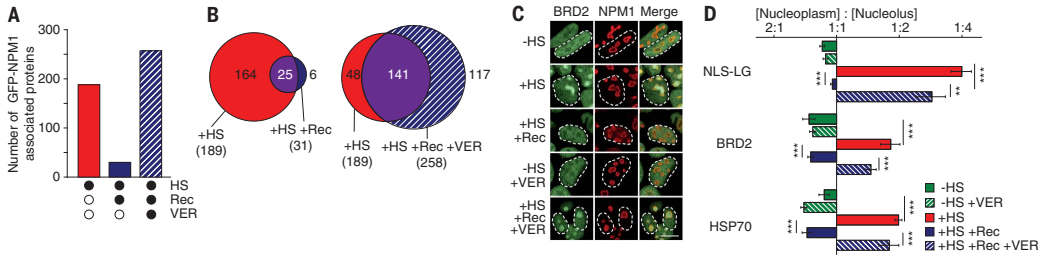


Fig. 2. GFP-NPM1 reversibly associates with endogenous proteins. (**A**) Number of GFP-NPM1–associated proteins (see table S1). GFP-NPM1 was transiently expressed in SILAC-labeled HEK293T cells before exposure to heat stress (+HS), followed by recovery (+HS +Rec) or recovery in the presence of Hsp70 inhibitor (+HS +Rec +VER). Control cells remained at 37°C (−HS). Anti-GFP immunoprecipitates from cell lysates were analyzed by mass spectrometry. Proteins that were enriched by a factor of ≥2 upon +HS over the −HS sample in at least three of four independent experiments were defined as being associated with GFP-NPM1 (see table S1). (**B**) Hsp70 inhibition prevents reversibility of GFP-NPM1 associations. Venn diagrams show distribution of GFP-NPM1–associated proteins under the conditions analyzed in (A). (**C**) Bromodomain-containing protein 2 (BRD2) reversibly accumulates in the nucleolus. HEK293T cells were treated as described above. Cells were immunostained for endogenous NPM1 and BRD2. Nuclei are marked by dashed circles. Representative images of three biological repeats are shown. Scale bar, 10 μm. (**D**) Partitioning of BRD2, NLS-LG, and Hsp70 between nucleoplasm and nucleoli in HEK293T cells treated as described above. Proteins were detected by immunostaining. Relative concentrations in nucleoplasm and nucleolus were quantified by measuring relative fluorescence intensities in 57 to 145 cells per condition across three biological repeats. $^{**}P \leq 0.05$, $^{***}P \leq 0.001$ (two-sided *t* test).

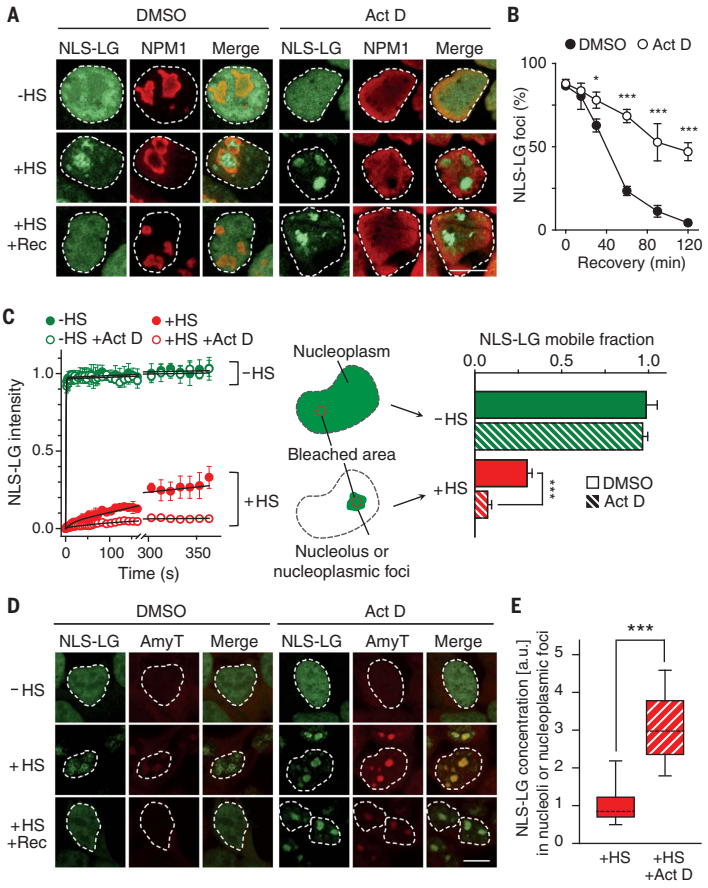


Fig. 3. The nucleolar environment prevents irreversible protein aggregation. (**A**) HEK293T cells expressing NLS-LG (green) were treated with actinomycin D (Act D) where indicated, followed by incubation with and without HS and recovery as in Fig. 1B. Cells were immunostained for NPM1 (red); nuclei are marked by dashed circles. (**B**) HEK293T cells expressing NLS-LG were treated as in (A) and recovery was monitored over 2 hours. Cells with nuclear NLS-LG foci were counted during recovery and expressed as percentage of total. Data are means ± SD; 453 to 693 cells were counted per time point and per condition across three biological repeats. *$P \leq 0.05$, ***$P \leq 0.001$ (two-sided t test). (**C**) HEK293T cells expressing NLS-LG were subjected to FRAP analysis. Cells were treated with Act D (open circles) before HS where indicated. For −HS experiments (green), the nucleoplasmic region was bleached, where NLS-LG localizes at 37°C. For +HS experiments (red), the nucleolus was bleached (see schematic). Left: Normalized FRAP curves with double-exponential fits. Curves represent means ± SD ($n \geq 3$ biological repeats). Right: Mobile fraction from the double-exponential fit. ***$P \leq 0.001$ (two-sided t test). (**D**) Cells expressing NLS-LG were subjected to Act D treatment where indicated, followed by heat stress (+HS) and stress with recovery (+HS +Rec), and stained with AmyT. Nuclei are marked by dashed circles. (**E**) Concentration of NLS-LG in the nucleolus and in nucleoplasmic aggregates (+Act D) upon heat stress. ***$P \leq 0.001$ (Mann-Whitney test; 100 measurements per condition across three biological repeats). Scale bars, 10 μm.

the high mobility of No-LG (Fig. 1, D and E, and fig. S3A). Thus, unfolding changed the interaction of luciferase with the GC phase. The larger hydrodynamic radius of unfolded luciferase may also contribute to the lower mobility. Consistently, the mobility of nucleolar GFP (No-GFP) (figs. S1C and S2C) remained unchanged upon heat stress (Fig. 1G).

HS also induced the formation of an immobile fraction of GFP-NPM1 (~30% of total) (Fig. 1F and fig. S3B), which returned to normal mobility upon recovery (fig. S3, B and C). Similar observations were made for nucleolin (GFP-NCL) (fig. S3, B and C). This suggested an association with unfolded (or misfolded) proteins that altered GC mobility. In support of this notion, expression of nucleolar luciferase (as a fusion with mScarlet; No-LS) further increased the immobile fraction of GFP-NPM1 upon HS (Fig. 1F and fig. S3B), which suggests that the amount of immobile GC protein correlated with the load of misfolded protein. In contrast, folded No-LS in control conditions had no effect on GFP-NPM1 mobility (Fig. 1F and fig. S3B). Indeed, endogenous NPM1 associated (directly or indirectly) with NLS-LG or No-LG upon HS by coimmunoprecipitation, but not in the absence of stress (fig. S3D). Thus, the unfolding of luciferase enhanced the association with the GC, consistent with a fraction of liquid-like NPM1 and nucleolin adopting a less dynamic state. Inhibiting Hsp70 activity completely inhibited the stress-denatured No-LG from recovery to normal mobility (Fig. 1E and fig. S3, A and E). Because No-LG remained localized to the nucleolus after refolding, this finding suggested that refolding mediated by Hsp70 was initiated in the nucleolus and was coupled with the mobilization of luciferase. Thus, upon proteotoxic stress, misfolded proteins immersed into the nucleolus, where they associated with GC proteins, thereby converting part of the liquid-like GC phase to a state of low mobility (Fig. 1F and fig. S3, B and C). Mobility was reestablished during recovery in an Hsp70-dependent manner, concomitant with refolding.

Endogenous proteins reversibly enter the nucleolus upon stress

To identify the endogenous proteins that enter the GC phase of the nucleolus upon stress, we performed GFP-NPM1 pull-down experiments followed by quantitative proteomics. We identified ~200 proteins that associated with NPM1 specifically upon HS, including numerous proteins of the nucleoplasm and nucleolus as well as some cytosolic proteins (Fig. 2A, fig. S4, A and B, and table S1). Thus, the stress-protective GC phase is accessible to proteins from both outside and within the nucleolus.

Nucleolin was also enriched in the NPM1 pull-down, but not the DFC marker fibrillarin (fig. S4C), suggesting an enhanced association between

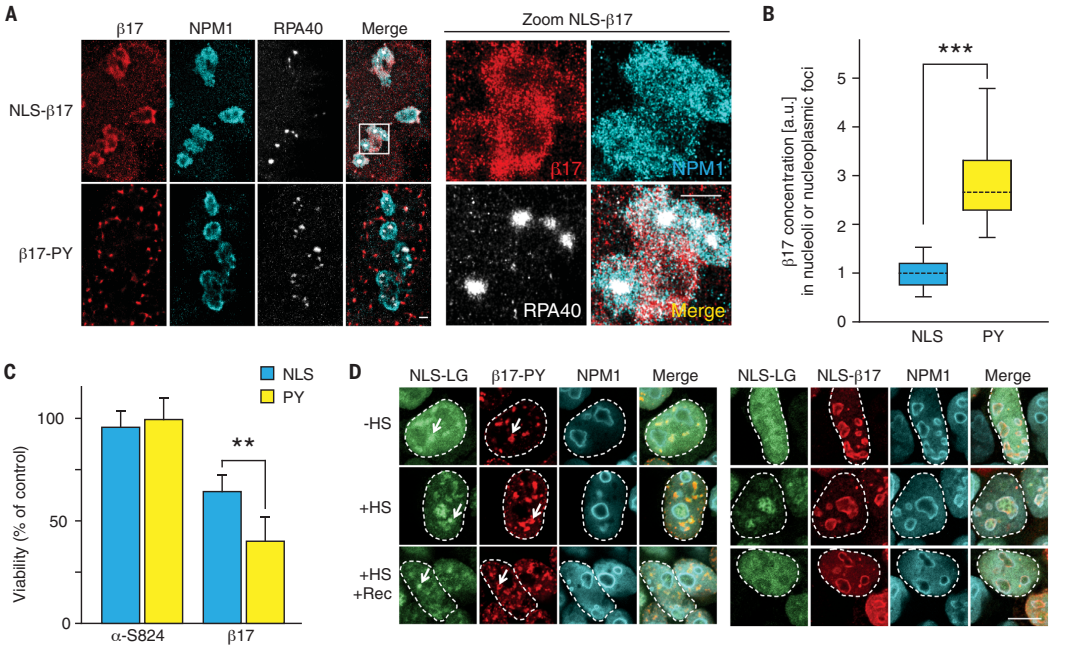


Fig. 4. Accumulation in the nucleolus reduces toxicity of amyloidogenic protein and prevents coaggregation with misfolded luciferase. (**A**) HEK293T cells were transfected with NLS-β17 or β17-PY prior to super-resolution imaging. Red, C-myc (β17); cyan, NPM1; white, RPA40. Zoomed images of NLS-β17 in the nucleolus are shown at right. (**B**) Density of β17 in the nucleolus (NLS-β17) and in nucleoplasmic aggregates (β17-PY) measured by superresolution imaging. Data were normalized to the average density of nucleolar NLS-β17. ***$P \leq 0.001$ (Mann-Whitney test). At least 36 and 52 measurements were performed on one representative experiment out of three biological repeats for NLS-β17 and β17-PY, respectively. (**C**) HEK293T cells were transfected with the indicated constructs and MTT cell viability assays were performed 4 days after transfection. Data were normalized to cells transfected with empty vector. Data are means + SD ($n \geq 3$). **$P \leq 0.01$ (two-sided t test). (**D**) β17-PY or NLS-β17 were transfected into the NLS-LG–expressing HEK293T cell line; 24 hours after transfection, cells were subjected to HS (+HS) and allowed to recover for 1 hour (+HS +Rec) before fixation. Cyan, endogenous NPM1; red, c-myc (β17). Arrows show NLS-LG sequestration into β17-PY aggregates. Scale bars, 1 μm (A), 10 μm (D).

NPM1 and nucleolin under heat stress, consistent with their reduced mobility (Fig. 1F and fig. S3, B and C). More than 400 proteins of the nucleoplasm or nucleolus were not enriched upon HS (fig. S4, A and C, and table S1). Thus, the proteins that entered the GC phase constituted a thermally sensitive subproteome. Indeed, these proteins were enriched in disordered and low-complexity sequences (fig. S4D), hallmarks of metastable structure. Their accumulation in the GC phase was reversible, whereas inhibition of Hsp70 preserved the association with NPM1 for most proteins (Fig. 2, A and B, fig. S5A, and tables S1 to S3). Additional proteins associated with NPM1 upon Hsp70 inhibition during recovery (Fig. 2, A and B, and tables S1 to S3).

We confirmed the reversible accumulation in the nucleolus for the proteins CDK1 and BRD2, which associated with NPM1 upon HS (Fig. 2, C and D, figs. S4C and S5, B and C, and table S1). A small but detectable fraction of total cellular Hsp70 also coprecipitated with NPM1 upon HS (fig. S5C), which suggests that associations with both Hsp70 and misfolded protein may contribute to forming the low-mobility GC fraction (Fig. 1, E and F, and fig. S3, B and C).

Functional relevance of the nucleolus in quality control

To explore the physiological importance of the nucleolus as a quality control compartment, we disrupted the nucleolar organization. Treating cells with a low concentration of the RNA polymerase I inhibitor actinomycin D (Act D) caused nucleolar disassembly and the release of NPM1 into the nucleoplasm (Fig. 3A) (*35*, *36*). NPM1 lost its liquid-like properties, as judged by its fast mobility (fig. S6A). NLS-LG was diffusely distributed in the nucleus of Act D–treated cells in the absence of stress but formed aggregate foci upon HS (Fig. 3A). These foci did not colocalize with NPM1. They resolved only slowly and inefficiently during recovery (Fig. 3, A and B) and sequestered Hsp70 for hours after the removal of stress (fig. S6B). The terminally misfolded CC* also formed persistent aggregates in Act D–treated cells, when proteasome function was inhibited (fig. S6, C and D). Thus, transport to the phase-separated GC compartment of the nucleolus was required to maintain misfolded proteins in a state competent for refolding or degradation once proteotoxic stress was relieved.

The NLS-LG in nucleoplasmic aggregates of Act D–treated cells was less mobile than nucleolar NLS-LG (Fig. 3C). Moreover, the nucleoplasmic foci were positive for amyloid (cross β structure)–specific dyes, in contrast to NLS-LG in the nucleolus (Fig. 3D and fig. S6E). Consistent with an amyloid-like state, the concentration of NLS-LG in nucleoplasmic foci was higher than in the nucleolus by a factor of ~3 (Fig. 3E). When nucleoli were disrupted, HS also caused endogenous proteins to form amyloid-like foci (fig. S6, F and G). Thus, entry of misfolded proteins into the nucleolus prevented amyloid-like aggregation.

We next analyzed the effect of the nucleolar environment on the model protein β17. This small β-sheet protein undergoes amyloidogenic aggregation and forms fibrils in vitro (*37*). Targeting

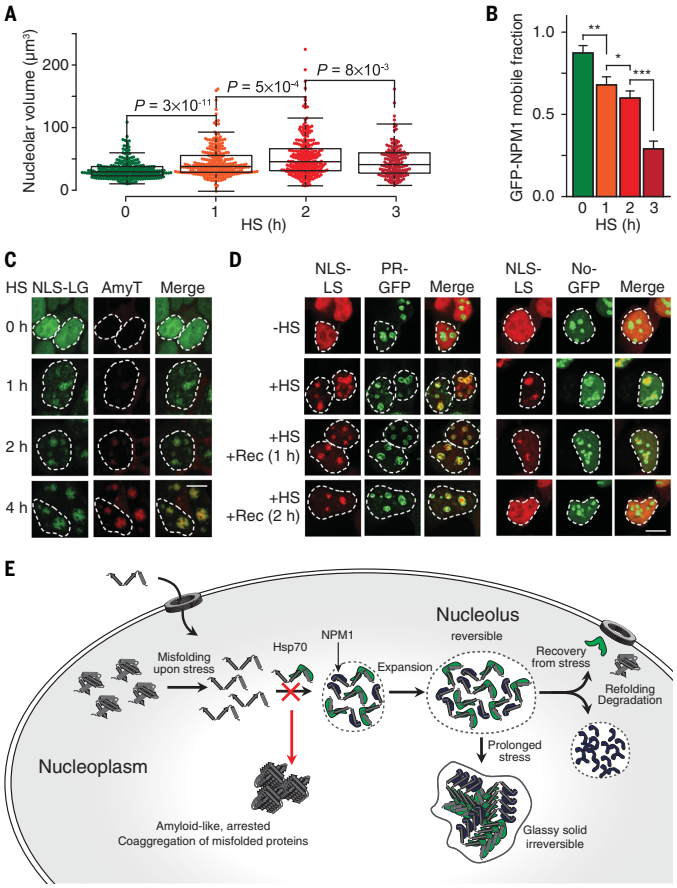


Fig. 5. The nucleolus changes phase properties during prolonged stress or accumulation of dipeptide repeat protein. (**A**) HeLa cells were incubated at 43°C for the number of hours indicated before staining for endogenous NPM1. The average nucleolar volume per nucleus is displayed as a bee-swarm box plot. Welch's *t* test was used to assess significant differences between conditions; the resulting *P* values are shown. Results are from three biological repeats representing 155 to 264 analyzed cells per condition. (**B**) NPM1 mobile fraction from FRAP experiments performed with HeLa cells transfected with GFP-NPM1. Cells were subjected to HS for the times indicated before and during FRAP measurement, and GFP-NPM1 mobile fractions were calculated. See also fig. S8, C and D. Data are means + SD of at least three biological repeats. $*P \le 0.05$, $**P \le 0.01$, $***P \le 0.001$ (two-sided *t* test). (**C**) NLS-LG–expressing HEK293T cells were subjected to heat treatment for the indicated times and stained with AmyT. Nuclei are marked by dashed circles. (**D**) HEK293T cells were cotransfected with NLS-LS and either PR_{175}-GFP or the nucleolar control protein No-GFP. Cells were maintained at 37°C (–HS) or subjected to heat stress (+HS) and recovery (+HS +Rec). (**E**) Model of nucleolar protein quality control for proteins entering the nucleolus from the nucleoplasm. Misfolded proteins immerse into the liquid-like GC phase of the nucleolus, presumably as a complex with Hsp70 (green), where they associate with GC proteins such as NPM1 (dark blue). There they are stored in an immobile state within the liquid-like GC phase, accompanied by an expansion of the nucleolus. Mobility is reestablished upon recovery from stress in an Hsp70-dependent manner, allowing refolding or proteasomal degradation in the nucleoplasm. Preventing access to the GC phase results in amyloid-like aggregation in the nucleoplasm. Upon prolonged stress, the GC phase increasingly transitions toward a more solid state. Misfolded proteins are no longer dispersed but form aggregates with amyloid-like properties. Scale bars, 10 µm.

β17 to the nucleus (NLS-β17) results in its accumulation in the nucleolus and a reduced toxicity relative to cytosolic β17 aggregates (*8*). To determine whether the nucleolar environment was responsible for this protective effect, we targeted β17 to the nucleoplasm by expressing it with the C-terminal nuclear localization signal PY (fig. S7A) (*38*). β17-PY formed foci in the nucleoplasm, whereas NLS-β17 accumulated in the GC phase of the nucleolus (Fig. 4A). Note that the NLS apparently functioned as a nucleolar targeting (or retention) signal in the sequence context with β17, but not in context with LG or GFP (Fig. 1B and fig. S2C). The function of the two localization sequences was position-independent (fig. S7A). The nucleoplasmic β17-PY aggregates were more concentrated than nucleolar NLS-β17 by a factor of 3 (Fig. 4B). β17-PY was more toxic than NLS-β17 (Fig. 4C), indicating that localization to the nucleolus reduced toxicity (*8*). The PY sequence per se did not confer toxicity (Fig. 4C and fig. S7B). As expected, nucleolar β17 variants but not nucleoplasmic β17-PY associated with NPM1 (fig. S7C). Moreover, NLS-β17-GFP was significantly more mobile than β17-GFP-PY (fig. S7, D and E), whereas disrupting the GC phase with Act D rendered NLS-β17-GFP less mobile (fig. S7, D and E).

Amyloid-like aggregates exert their toxic effect in part by coaggregating and sequestering essential, metastable proteins (*8*, *39*–*41*). Indeed, the nucleoplasmic β17-PY aggregates sequestered NLS-LG upon HS, thereby preventing NLS-LG from entering the nucleolus (Fig. 4D). Nucleolar NLS-β17 had no such effect and did not prevent repartitioning of NLS-LG to the nucleoplasm upon recovery (Fig. 4D). Thus, the GC phase of the nucleolus has the capacity to simultaneously store different proteins and allow them to undergo selective renaturation.

Accumulation of misfolded proteins in the nucleolus did not interfere with ribosome biogenesis, as nucleolar NLS-β17 did not interfere with the assembly and export of yellow fluorescent protein (YFP)–tagged 40*S* ribosomal protein S2 (RPS2-YFP) to the cytosol (fig. S7, F and G) (*42*). In contrast, nucleoplasmic aggregates of β17-PY caused coaggregation of RPS2-YFP and nuclear retention (fig. S7, F and G).

Limitations of nucleolar quality control

To explore the capacity of the nucleolus for incorporating misfolded proteins, we exposed cells to prolonged stress. We observed a significant increase in nucleolar volume during the first 2 hours of HS (Fig. 5A), presumably reflecting the influx of misfolded proteins. The nucleoli lost their liquid droplet–like appearance and adopted irregular shapes (fig. S8, A and B), suggestive of a transition to a hardened state. Indeed, the mobile fraction of GFP-NPM1 decreased markedly during prolonged HS (Fig. 5B and fig. S8, C and D). To further assess these changes, we stained NLS-LG–expressing cells with the amyloid-specific dye AmyT and observed a distinct nucleolar staining that developed over time (Fig. 5C). The foci that formed during extended HS dissolved only slowly upon recovery (fig. S8E). Apparently, prolonged stress exhausted the storage

capacity of the nucleolus for misfolded proteins, resulting in a transition to a solid, aggregated state.

Expression of C9orf72 encoded dipeptide repeat proteins (DPRs) is a possible cause of familial ALS and frontotemporal dementia (FTD) (*43*–*45*). These peptides cause nucleolar dysfunction by modulating the liquid-like properties of the nucleolus (*19*, *20*). We expressed the DPR-protein PR_{175}-GFP along with nuclear luciferase (NLS-LS). PR_{175}-GFP incorporated efficiently into the GC phase of the nucleolus (Fig. 5D) (*19*, *20*), resulting in reduced mobility of a fraction of mScarlet-NPM1 (fig. S9A). NLS-LS entered the nucleolus during HS and colocalized with PR_{175}-GFP but failed to repartition during recovery (Fig. 5D), remaining arrested in the nucleolus for hours (fig. S9B). In contrast, control cells expressing No-GFP allowed normal NLS-LS repartitioning (Fig. 5D and fig. S9B). Thus, nucleolar DPR protein leads to a breakdown of nucleolar quality control, which may contribute to the cellular pathology in ALS and FTD.

Conclusions

The liquid-like GC phase of the nucleolus functions as a non–membrane-bound protein quality control compartment (Fig. 5E). It is characterized by a remarkable chaperone-like capacity to prevent irreversible aggregation of misfolded proteins, facilitating refolding during recovery from stress. Misfolded proteins associate with nucleolar proteins including NPM1, thereby converting a fraction of the GC phase to a less dynamic state (Fig. 5E). The association of misfolded proteins with the GC phase is regulated by the chaperone Hsp70, which is required for refolding (Fig. 5E). Nucleoplasmic proteins exit the nucleolus upon refolding, and nucleolar proteins resume their functional state. However, the capacity of the nucleolus to store misfolded proteins is limited, and prolonged stress causes aberrant phase behavior associated with the danger of irreversible aggregation (Fig. 5E). Moreover, disease-related DPR proteins impair the ability of the nucleolus to reversibly store misfolded proteins—a mechanism that may contribute to neurodegenerative pathology.

REFERENCES AND NOTES

1. W. E. Balch, R. I. Morimoto, A. Dillin, J. W. Kelly, *Science* **319**, 916–919 (2008).
2. T. Gidalevitz, E. A. Kikis, R. I. Morimoto, *Curr. Opin. Struct. Biol.* **20**, 23–32 (2010).
3. E. M. Sontag, R. S. Samant, J. Frydman, *Annu. Rev. Biochem.* **86**, 97–122 (2017).
4. R. Higuchi-Sanabria, P. A. Frankino, J. W. Paul 3rd, S. U. Tronnes, A. Dillin, *Dev. Cell* **44**, 139–163 (2018).
5. T. Shpilka, C. M. Haynes, *Nat. Rev. Mol. Cell Biol.* **19**, 109–120 (2018).
6. P. Walter, D. Ron, *Science* **334**, 1081–1086 (2011).
7. A. Korennykh, P. Walter, *Annu. Rev. Cell Dev. Biol.* **28**, 251–277 (2012).
8. A. C. Woerner *et al*., *Science* **351**, 173–176 (2016).
9. L. Vincenz-Donnelly *et al*., *EMBO J.* **37**, 337–350 (2018).
10. E. Rousseau *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9648–9653 (2004).
11. J. S. Andersen *et al*., *Nature* **433**, 77–83 (2005).
12. Z. M. March, O. D. King, J. Shorter, *Brain Res.* **1647**, 9–18 (2016).
13. A. J. Baldwin *et al*., *J. Am. Chem. Soc.* **133**, 14160–14163 (2011).
14. S. Raychaudhuri *et al*., *Cell* **156**, 975–985 (2014).
15. R. D. Jones, R. G. Gardner, *Curr. Opin. Cell Biol.* **40**, 81–89 (2016).
16. A. von Mikecz, *Nucleus* **5**, 311–317 (2014).
17. C. G. Chung, H. Lee, S. B. Lee, *Cell. Mol. Life Sci.* **75**, 3159–3180 (2018).
18. K. Mori *et al*., *Science* **339**, 1335–1338 (2013).
19. I. Kwon *et al*., *Science* **345**, 1139–1145 (2014).
20. K. H. Lee *et al*., *Cell* **167**, 774–788.e17 (2016).
21. J. E. Sleeman, L. Trinkle-Mulcahy, *Curr. Opin. Cell Biol.* **28**, 76–83 (2014).
22. M. Feric *et al*., *Cell* **165**, 1686–1697 (2016).
23. M. Biggiogera *et al*., *Development* **110**, 1263–1270 (1990).
24. D. M. Mitrea *et al*., *Nat. Commun.* **9**, 842 (2018).
25. J. K. Box *et al*., *BMC Mol. Biol.* **17**, 19 (2016).
26. J. M. Velazquez, S. Lindquist, *Cell* **36**, 655–662 (1984).
27. W. J. Welch, J. R. Feramisco, *J. Biol. Chem.* **259**, 4501–4513 (1984).
28. H. Pelham, M. Lewis, S. Lindquist, *Philos. Trans. R. Soc. London Ser. B* **307**, 301–307 (1984).
29. E. A. Nollen *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12038–12043 (2001).
30. J. Schnitzbauer, M. T. Strauss, T. Schlichthaerle, F. Schueder, R. Jungmann, *Nat. Protoc.* **12**, 1198–1228 (2017).
31. S. Strauss *et al*., *Nat. Methods* **15**, 685–688 (2018).
32. R. Schlecht *et al*., *PLOS ONE* **8**, e78443 (2013).
33. S. H. Park *et al*., *Cell* **154**, 134–145 (2013).
34. A. Birbach, S. T. Bailey, S. Ghosh, J. A. Schmid, *J. Cell Sci.* **117**, 3615–3624 (2004).
35. M. Chen, P. Jiang, *Acta Pharmacol. Sin.* **25**, 902–906 (2004).
36. T. Dousset *et al*., *Mol. Biol. Cell* **11**, 2705–2717 (2000).
37. M. W. West *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11211–11216 (1999).
38. J. Gal *et al*., *Neurobiol. Aging* **32**, 2323.e27–2323.e40 (2011).
39. H. Olzscha *et al*., *Cell* **144**, 67–78 (2011).
40. M. S. Hipp, S. H. Park, F. U. Hartl, *Trends Cell Biol.* **24**, 506–514 (2014).
41. Y. J. Zhang *et al*., *Nat. Neurosci.* **19**, 668–677 (2016).
42. T. Wild *et al*., *PLOS Biol.* **8**, e1000522 (2010).
43. M. DeJesus-Hernandez *et al*., *Neuron* **72**, 245–256 (2011).
44. I. Gijselinck *et al*., *Lancet Neurol.* **11**, 54–65 (2012).
45. A. E. Renton *et al*., *Neuron* **72**, 257–268 (2011).

ACKNOWLEDGMENTS

We thank U. Kutay for the RPS2-YFP HeLa cell line; D. Edbauer for the expression plasmid PR_{175}-GFP; B. Sperl, O. K. Wade, and S. Strauss for technical assistance; and A. Ries for support with SILAC-MS/MS. We acknowledge support by the MPIB Imaging facility and G. Cardone for providing the algorithm for image quantification. **Funding:** F.F. was supported by an EMBO Long Term Fellowship. The research leading to these results has received funding from the European Commission under grant FP7 GA ERC-2012-SyG_318987–ToPAG, and MolMap grant agreement 680241, the Munich Cluster for Systems Neurology, and the Max Planck Foundation. **Author contributions:** F.F. designed and performed most of the experiments. R.G. carried out initial experiments. F.S. and T.S. carried out the high resolution imaging. R.K. supervised the proteomic analysis and S.T. and J.C. analyzed sequence properties of NPM1 associated proteins. R.J. designed and supervised the high-resolution imaging experiments. F.U.H. and M.S.H. initiated and supervised the project and wrote the paper with input from F.F. and the other authors. **Competing interests:** J.C. is also affiliated with the Department of Biological and Medical Psychology, Faculty of Psychology, University of Bergen, Bergen, Norway. The authors declare no other competing interests. **Data and materials availability:** Data from the mass spectrometry analysis described in this manuscript can be found in the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/365/6451/342/suppl/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S3
References (*46*–*61*)

5 February 2019; resubmitted 23 May 2019
Accepted 27 June 2019
Published online 11 July 2019
10.1126/science.aaw9157

2.4 Complete submission in PRIDE database

The PRoteomics IDEntifications (PRIDE) repository is one of the largest proteomics data repositories worldwide. PRIDE data along with Ensembl, UniProt, and Expression Atlas are becoming very valuable resources in proteomics. The PRIDE database https://www.ebi.ac.uk/pride/ was set up in 2004 at the European Bioinformatics Institute (EMBL-EBI, Hinxton, Cambridge, UK) to enable public data deposition of mass spectrometry proteomics data, providing access to the experimental data described in scientific publications. In this article, the authors summarize the developments in PRIDE resources and related tools. Proteomics data standard tab-delimited mzTab format[125], developed by the Proteomics Standards Initiative (PSI) was added to do complete submission in the PRIDE repository. In complete submission along with peptide/protein identifications, also the corresponding quantitative information, are parsed and linked to the originating mass spectra. In Maxquant, mzTab output table that is required for the complete submission was missing. Dr. Şule Yılmaz from Cox Lab and I developed the mzTab output table in MaxQuant. The mzTab output table contains metadata, protein, peptide, PSM, and small molecule sections in table based format containing their basic information. We provide metadata, protein and PSM section in mzTab output file. The peptide section, which is recommended for quantitative information at peptide level will be implemented in the future. Users have to check the mzTab box while running the MaxQuant software and use the table for the complete submission to the PRIDE repository.

Yasset Perez-Riverol, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J. Kundu, Avinash Inuganti, Johannes Griss, Gerhard Mayer, Martin Eisenacher, Enrique Pérez, Julian Uszkoreit, Julianus Pfeuffer, Timo Sachsenberg, Şule Yılmaz, **Shivani Tiwary**, Jürgen Cox, Enrique Audain, Mathias Walzer, Andrew F. Jarnuczak, Tobias Ternent, Alvis Brazma, and Juan Antonio Vizcaíno. The pride database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*, 47(D1):D442–D450, nov 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1106

D442–D450 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 5 November 2018*
doi: 10.1093/nar/gky1106

The PRIDE database and related tools and resources in 2019: improving support for quantification data

Yasset Perez-Riverol[1,*], Attila Csordas[1], Jingwen Bai[1], Manuel Bernal-Llinares[1], Suresh Hewapathirana[1], Deepti J. Kundu[1], Avinash Inuganti[1], Johannes Griss[1,2], Gerhard Mayer[3], Martin Eisenacher[3], Enrique Pérez[1], Julian Uszkoreit[3], Julianus Pfeuffer[4], Timo Sachsenberg[4], Şule Yılmaz[5], Shivani Tiwary[5], Jürgen Cox[5], Enrique Audain[6], Mathias Walzer[1], Andrew F. Jarnuczak[1], Tobias Ternent[1], Alvis Brazma[1] and Juan Antonio Vizcaíno[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Division of Immunology, Allergy and Infectious Diseases, Department of Dermatology, Medical University of Vienna, Vienna, 1090, Austria, [3]Ruhr University Bochum, Medical Faculty, Medizinisches Proteom-Center, D-44801 Bochum, Germany, [4]Applied Bioinformatics, Department for Computer Science, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany, [5]Computational Systems Biochemistry, Max Planck Institute for Biochemistry, Martinsried, 82152, Germany and [6]Department of Congenital Heart Disease and Pediatric Cardiology, Universitätsklinikum Schleswig–Holstein Kiel, Kiel, 24105, Germany

Received September 22, 2018; Revised October 19, 2018; Editorial Decision October 22, 2018; Accepted October 22, 2018

ABSTRACT

The PRoteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/) is the world's largest data repository of mass spectrometry-based proteomics data, and is one of the founding members of the global ProteomeXchange (PX) consortium. In this manuscript, we summarize the developments in PRIDE resources and related tools since the previous update manuscript was published in *Nucleic Acids Research* in 2016. In the last 3 years, public data sharing through PRIDE (as part of PX) has definitely become the norm in the field. In parallel, data re-use of public proteomics data has increased enormously, with multiple applications. We first describe the new architecture of PRIDE Archive, the archival component of PRIDE. PRIDE Archive and the related data submission framework have been further developed to support the increase in submitted data volumes and additional data types. A new scalable and fault tolerant storage backend, Application Programming Interface and web interface have been implemented, as a part of an ongoing process. Additionally, we emphasize the improved support for quantitative proteomics data through the mzTab format. At last, we outline key statistics on the current data contents and volume of downloads, and how PRIDE data are starting to be disseminated to added-value resources including Ensembl, UniProt and Expression Atlas.

INTRODUCTION

High-throughput mass spectrometry (MS)-based proteomics approaches have matured significantly in recent years, becoming an increasingly used tool in biological research, sometimes together with other 'omics' approaches such as genomics and transcriptomics. Similarly, to what happened in those fields, in the last 15 years several public proteomics repositories and bioinformatics resources have been developed to support proteomics researchers. The PRoteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/) was set up in 2004 at the European Bioinformatics Institute (EMBL-EBI, Hinxton, Cambridge, UK) to enable public data deposition of MS proteomics data, providing access to the experimental data described in scientific publications (1). Since then, PRIDE (more concretely its archival component, PRIDE Archive) has evolved in parallel with the field becoming the largest proteomics data repository worldwide (2).

Although datasets coming from data-dependent acquisition (DDA) proteomics approaches represent by far the most abundant type of experiment, PRIDE Archive can

*To whom correspondence should be addressed. Tel: +44 0 1223 492513; Fax: 01223 484696; Email: yperez@ebi.ac.uk
Correspondence may also be addressed to Dr. Juan Antonio Vizcaíno. Tel: +44 0 1223 492686; Fax: 01223 484696; Email: juan@ebi.ac.uk

Nucleic Acids Research, 2019, Vol. 47, Database issue **D443**

store datasets coming from all main proteomics data workflows (including Data Independent Acquisition (DIA), MS imaging, and top down proteomics, among others). The mandatory data types to be included in each submitted dataset are the raw files (output files from the mass spectrometers) and the processed results (at least peptide/protein identification results, quantification information is optional). Therefore, each dataset in PRIDE Archive can contain heterogeneous data types such as peptide/protein identifications and quantification values, the mass spectra (peak lists and raw data), the searched sequence databases or spectral libraries, programming scripts and any other technical and/or biological metadata provided by the data submitters.

A key development led by PRIDE was the establishment of the ProteomeXchange (PX) consortium of MS proteomics resources (http://www.proteomexchange.org) (3), with the overall aim of standardizing data submission and dissemination of proteomics data worldwide. By September 2018, the following proteomics resources are also part of PX: PeptideAtlas and PASSEL (PeptideAtlaS SRM Experiment Library) (4,5), MassIVE (http://massive.ucsd.edu/), jPOSTrepo (6), iProx (http://www.iprox.org/) and Panorama Public (7).

PRIDE has four major aims: (i) support data deposition of proteomics experiments, and perform automatic and manual curation of the related experimental metadata; (ii) implement quality control pipelines and visualization components to enable the assessment of the data quality (8); (iii) promote and facilitate the re-use of public proteomics data; and at last, (iv) disseminate high-quality proteomics evidences to added-value resources, including Ensembl (9), UniProt (10) and Expression Atlas (11).

In order to facilitate the deposition, visualization and quality assessment of the data, the team has developed over the years a complete framework of open-source software, including stand-alone tools such as the PX Submission tool and PRIDE Inspector (12). In addition, the different PRIDE related data pipelines, REST web services (13) and the web interfaces (2) have been continuously refined. Furthermore, we have developed a number of open source software libraries in Java, including jmzML, jmzIdentML, jmzReader, jmzTab, ms-data-core-api (14) and the PIA (Protein Inference Algorithms) toolbox (15,16) (https://github.com/PRIDE-Utilities), to support handling (e.g. read and writing) of the most popular proteomics data standard formats (e.g. mzML, mzIdentML, mzTab) developed by the Proteomics Standards Initiative (PSI) (17). In addition to all the PX resources mentioned above, there are additional proteomics databases and resources available providing protein expression information, most notably the Global Proteome Machine Database (GPMDB) (18), the CPTAC (Clinical Proteomic Tumor Analysis Consortium) data portal (19) and ProteomicsDB (20).

In this manuscript, we will summarize the main PRIDE related developments in the last three years, since the previous *Nucleic Acids Research* database update manuscript was published (2). We will discuss PRIDE Archive in more detail but will also provide updated information about the PRIDE related tools and other ongoing activities.

CURRENT STATUS OF PRIDE ARCHIVE AND RELATED TOOLS

Original submitted datasets by scientists are stored in PRIDE Archive (http://www.ebi.ac.uk/pride/archive/). All datasets remain private (password protected) by default and are only made publicly available after the related manuscript has been accepted, or when PRIDE is notified to do so by the original submitter. Data in PRIDE Archive can be searched and accessed in four different ways: (i) the web interface, providing a general overview of each dataset; (ii) the PRIDE Inspector tool (12), which can be used for downloading the submitted data files and to visualize spectrum, peptide and protein information in open formats, including several PSI standards; (iii) the Restful web service (https://www.ebi.ac.uk/pride/ws/archive/) (21); and (iv) a file repository, where both the FTP and Aspera (https://asperasoft.com/) file transfer protocols can be used to access the files. In addition, all public datasets in PRIDE Archive are available through OmicsDI (https://www.omicsdi.org/), an EMBL-EBI resource which integrates public datasets coming from different omics technologies (22). Figure 1 provides an overview of the PRIDE ecosystem, including the most relevant tools, software libraries and the data dissemination into other resources.

New PRIDE Archive infrastructure: scaling-up a resource for present-day proteomics experiments

The number of datasets submitted to PRIDE has grown very significantly in recent years, in parallel with the size of the experiments, e.g. the number of samples, biological/technical replicates and evidences—mass spectra, Peptide Spectrum Matches (PSMs), peptides and proteins. Two different factors, scalability and reliability (fault-tolerance), have guided the development of the new PRIDE Archive distributed architecture (Supplementary Note 1). Every storage item (e.g. MongoDB, Solr Indexes) is now deployed in two EMBL-EBI datacenters as shard distributed clusters. This new architecture ensures that if one datacenter is not accessible (due to e.g. technical maintenance), PRIDE Archive is still accessible.

Data submission process: improved support for quantification results

The data submission process has not substantially changed because the overall PX submission guidelines have remained stable (23). An updated web tutorial explaining the process is available at http://www.ebi.ac.uk/training/online/course/proteomexchange-submissions-pride. The main addition is the support for the standardized tab-delimited mzTab format (24) to perform 'Complete' submissions (those where peptide/protein identifications and, thanks to this ongoing development, also the corresponding quantitative information, can be parsed by the repository, made accessible in the database and linked to the originating mass spectra). Therefore, support for mzTab has enabled the deposition of quantitative data into PRIDE Archive for the first time in a standard format that is supported for 'Complete' submissions (Supplementary Note 2). By

D444 Nucleic Acids Research, 2019, Vol. 47, Database issue

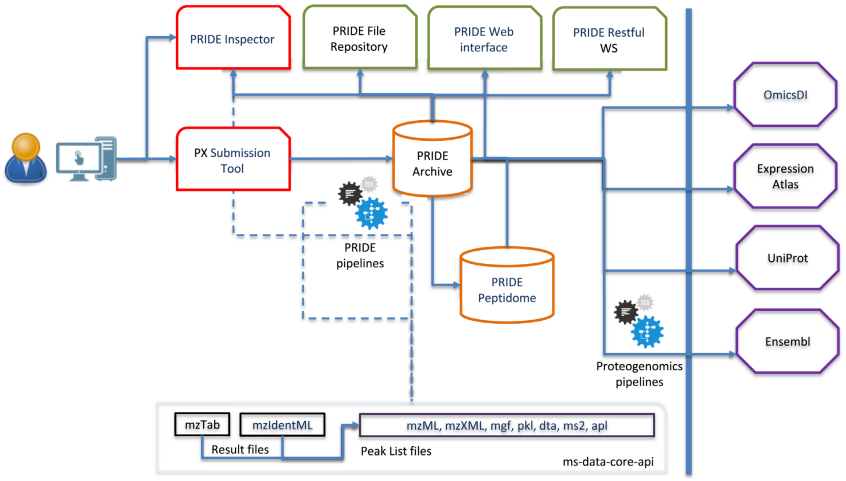


Figure 1. Overview of the PRIDE ecosystem, including the resources (PRIDE Archive and PRIDE Peptidome, in orange), tools (PRIDE Inspector and PX Submission Tool, in red), software libraries (in black), web interface and API (in green) and the external resources where PRIDE data are disseminated to (in purple).

October 2018, the Mascot (25) search engine (e.g. https://www.ebi.ac.uk/pride/archive/projects/PXD009079), the OpenMS framework (26) (e.g. https://www.ebi.ac.uk/pride/archive/projects/PXD010981) and MaxQuant (27) (e.g. https://www.ebi.ac.uk/pride/archive/projects/PXD011194) enable natively the export of quantitative results into mzTab. In order to keep improving the support for quantification data, we aim to promote the implementation of mzTab in other popular software tools such as Proteome Discoverer (*ThermoFisher Scientific*).

The mzIdentML format remains as the mainstream format for 'Complete' submissions and is increasingly supported by search engines and tools (14). In case mzTab and/or mzIdentML are not yet supported by the user's software of choice, the alternative is to perform a 'Partial' submission, which is also the current alternative for data workflows such as DIA, top-down and MS imaging. In parallel with the ongoing developments in PSI data standard formats, all PRIDE-related software libraries (https://github.com/PRIDE-Utilities) have been continuously developed, making data handling and submission a much more robust process. In this context, we will continue extending our libraries (ms-data-core-api and jmzIdentML) to support the new features included in mzIdentML version 1.2 such as MS/MS cross-linking and proteogenomics approaches.

The PX submission tool

The PX Submission tool (3) (available at https://github.com/proteomexchange/px-submission-tool) is a stand-alone tool used by most PRIDE submitters to perform data submissions. Some of the recent refinements done in the tool are: (i) the integration of the new OLS (Ontology Lookup Service) Client and OLS Dialog libraries (28), supporting the new version of the OLS, used to annotate datasets using controlled vocabulary terms; and (ii) the addition of a direct feedback system for users to report how the data submission went.

PRIDE Inspector toolsuite: reviewing datasets before and after submission to PRIDE Archive

The PRIDE Inspector tool (12) (available at https://github.com/PRIDE-Toolsuite/pride-inspector) was developed to enable researchers to visualize and perform an initial quality assessment of the data both before and after data submissions are performed, once the dataset becomes public. PRIDE Inspector supports several different experimental open output files, ranging from mass spectra (mzML, mzXML and the most popular peak lists formats such as mgf, dta, ms2, pkl and apl), identification results (mzIdentML, mzTab), to quantification data (mzTab). Some refinements have been implemented in the tool and in the underlying software libraries over these last years. The main new feature added recently is the support for reviewers to download private datasets using the much faster Aspera file transfer protocol. This key functionality to facilitate the review process is not available via the PRIDE Archive web interface at present.

PRIDE web interface and restful API: retrieving public proteomics data

The PRIDE web interface and Restful API (13) can be used to retrieve and visualize the data corresponding to all PRIDE datasets. The new PRIDE web interface (Figure 2) provides a powerful mechanism to search and/or filter by several types of metadata information, such as sample

Nucleic Acids Research, 2019, Vol. 47, Database issue ***D445***

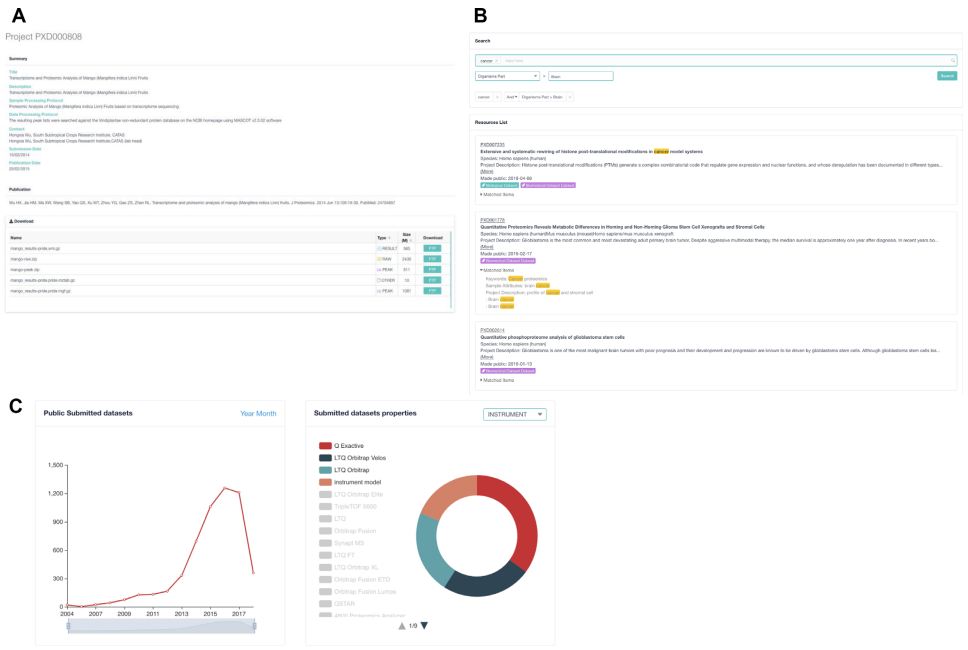


Figure 2. Screenshots of the new PRIDE Archive web interface. (**A**) The project (dataset) page provides a general overview of every submitted dataset. (**B**) The PRIDE Archive search page, where it is possible for users to query PRIDE Archive using keywords and additional properties such as species, tissues and instruments, among others. (**C**) Real-time statistics (including number of submitted datasets per month, number of submitted datasets per instrument type, etc.) are now provided.

details (e.g. species, tissue, cell type, etc.), instrumentation (mass spectrometer), keywords and other provided annotations (Supplementary Note 3). Using the API, it is possible to programmatically query for and retrieve peptide and protein identifications, dataset and assay specific metadata, and all the originally submitted files. Both components are currently under development and new functionalities are being implemented such as suggestions for similar datasets, autocomplete search capabilities and live data content statistics (Figure 2).

PRIDE Peptidome: high-quality peptide evidence from PRIDE Archive

The PSMs reported in PRIDE Archive are quality-filtered using a spectrum clustering approach (29). All the identified spectra coming from the public experiments in PRIDE Archive were clustered using the second iteration of the PRIDE Cluster algorithm, called *spectra-cluster* (https://github.com/spectra-cluster) (30). The results of the clustering process are made available through the peptide centric PRIDE Peptidome resource (formally known as PRIDE Cluster, http://www.ebi.ac.uk/pride/cluster/), which has a completely new web interface, in line with the new PRIDE Archive one. The corresponding spectral libraries and spectral archives (containing only unidentified spectra) are made available at https://www.ebi.ac.uk/pride/cluster/#/libraries and at https://www.ebi.ac.uk/pride/cluster/#/results.

PRIDE ARCHIVE DATA CONTENT STATISTICS

By 1 September 2018, PRIDE Archive contained 10100 datasets (compared to 3336 datasets on September 2015), of which roughly 19% are 'Complete' (1975 datasets), 72% are 'Partial' (7295) and the remaining 9% (830) correspond to old 'legacy' datasets submitted before the PX data workflow was implemented. Figure 3A shows the evolution in the number of submitted datasets per month. By September 2018, an average of 274 datasets were submitted per month during 2018, amounting to more than 2-fold when compared with 3 years ago. The landmark dataset PXD010000 was submitted on 1 June 2018. These figures correspond to all datasets including private ones (non-released, password protected). By 1 September 2018, 56% (5719) of the datasets were publicly available. Interestingly, the number of submitted datasets generated using experimental approaches other than DDA is growing (Figure 3B). By September 2018, the number of datasets classified as DDA in PRIDE was 91%, while 26% was classified as other types (Figure 3B). The number of DIA and selected reaction monitoring (SRM)

D446 Nucleic Acids Research, 2019, Vol. 47, Database issue

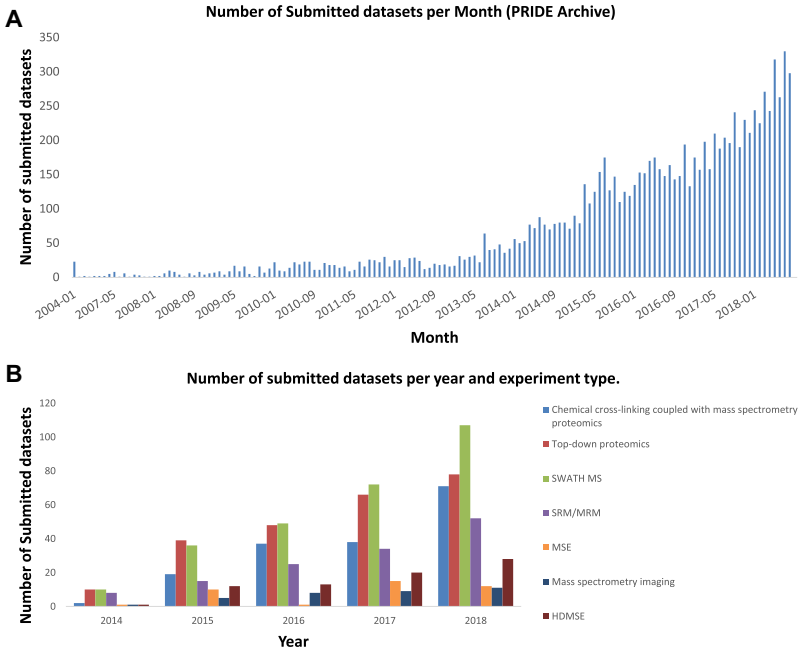


Figure 3. (**A**) Number of submitted datasets to PRIDE per month (from beginning of 2004 till September 2018). (**B**) Number of submitted datasets per experimental approach per year (from 2014 till September 2018).

datasets are indeed the most abundant ones behind DDA datasets.

The most represented species (including both public and private datasets) are human (4335 datasets) and some of the main model organisms, most notably mouse (1432), *Arabidopsis thaliana* (375), *Saccharomyces cerevisiae* (341), rat (300), *Escherichia coli* (247), cow (112), *Drosophila melanogaster* (101), chicken (65), rice (70) and soybean (49). Overall, datasets coming from more than 1840 different taxonomy identifiers are stored in PRIDE Archive (Figure 4). These statistics represent in our view a fair reflection of the current guidelines for mandatory data deposition developed by many funding agencies and some scientific journals. At the time of writing, the Wellcome Trust, BBSRC, MRC and the NIH, among other funders, mandate or strongly encourage open access to research data including proteomics. Additionally, two of the most prominent proteomics journals (*Molecular and Cellular Proteomics* and *Journal of Proteome Research*) and journals from the *Nature* group now mandate submission of at least the raw data supporting each proteomics publication. Other journals already recommend or strongly recommend data submission (e.g. *Proteomics* (Wiley), *Journal of Proteomics* (Elsevier), *PLOS* journals, etc.). The evolution in the percentage of research articles supported by PRIDE datasets (in three different proteomics journals: *Molecular and Cellular Proteomics, Journal of Proteome Research* and *Proteomics*) is explained in Supplementary Note 4. At last, in this context, it is important to highlight that the Human Proteome Project has developed formal guidelines mandating data submission for all generated datasets (31).

DATA RE-USE OF PUBLIC PRIDE DATASETS

Proteomics researchers are increasingly re-using public data available in PRIDE (and other resources) for a broad range of purposes. We came up with four categories of public proteomics data re-use: (i) *use*, (ii) *re-use*, (iii) *reprocess* and (iv) *repurpose* (32). A simple example of the direct *use* of data are given by connecting information between proteomics data resources and other resources such as UniProt and Ensembl (10). In the case of *re-use*, public data are re-used in novel experiments with the potential of generating new knowledge. Data from a large number of independent datasets can be analyzed or re-used in combination (a so-called *meta*-analysis study), to extract new knowledge not accessible from any individual dataset. In the case of *reprocess*, public datasets are re-analyzed to provide an updated or integrated view on the results, as protein sequence databases and software tools evolve. At last, *repurposing* includes all those cases where the data are considered in a context that is different to that of the original experiment. Two popular applications are proteogenomics approaches (for human and the main model organisms, e.g. (33,34)), and the discov-

Nucleic Acids Research, 2019, Vol. 47, Database issue ***D447***

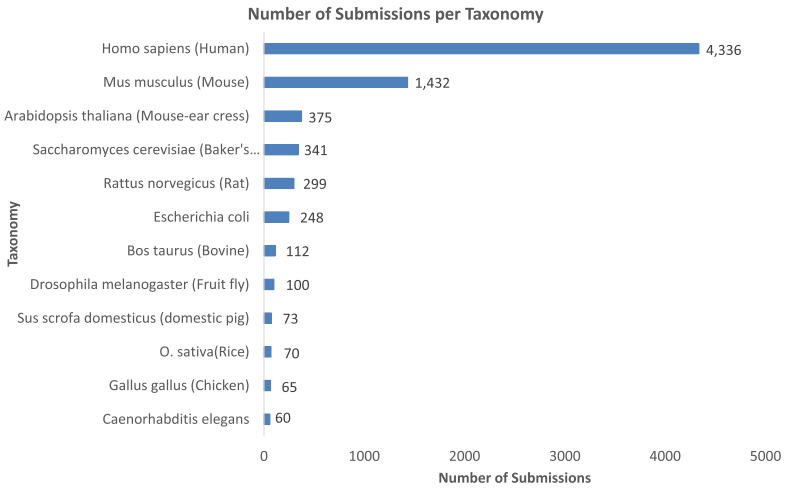


Figure 4. Number of submitted datasets to PRIDE Archive per taxonomy identifier.

ery of novel PTMs (Post-Translational Modifications). Recent reviews in re-use of public proteomics data are available (32,35).

To corroborate the increase in data re-use, Figure 5 shows the increase in the volume of PRIDE data downloads per year, reaching 296 TBs during 2017. In addition, using the previously mentioned resource OmicsDI, it is now possible to trace the number of re-analyses of PRIDE datasets performed by PeptideAtlas and GPMDB and the number of direct citations of PRIDE datasets in the literature (BioRxiv: https://doi.org/10.1101/282517). By September 2018, 293 datasets had been re-analyzed and 381 dataset identifiers had been cited directly in the literature.

PRIDE Proteogenomics: representing peptide sequences into Ensembl using 'TrackHubs'

The PRIDE and Ensembl teams have been working together to improve the integration of proteomics data in a genome context. Peptide evidence from 'complete' public datasets in PRIDE Archive are first quality-filtered (at a 1% peptide false discovery rate) using a framework that uses PIA (15). Reliable peptide sequences (including PTMs) are mapped to the corresponding genomic coordinates from a given Ensembl release using the PoGo tool (36). The resulting data for each individual dataset is then combined and made available through the Ensembl 'TrackHub' registry, using the popular BED format. In addition to individual datasets, PRIDE Cluster data (now re-named to PRIDE Peptidome) is also made available as independent 'TrackHubs'. At the time of writing, 184 PRIDE public datasets have been already made available in the Ensembl 'TrackHub' registry (https://www.trackhubregistry.org/): 163 human, 15 from *Mus musculus*, 4 from *Rattus norvegicus* and 2 from *Bos taurus*. The 'TrackHubs' can be searched in the 'TrackHub' registry by project identifier, taxonomy and/or specific keywords available in the description of the corresponding PRIDE dataset. As a key point, 'TrackHubs' can be loaded and visualized in the Ensembl web interface together with other genomic features (Figure 6). More than 4 million peptide sequences (1.2 million of them containing PTMs) have been mapped to the human genome (GRCh38). We are working in including data coming from other model organisms. It is very important to highlight that the developed framework supports the other two major genome browsers: The UCSC Genome Browser and IGV (Integrative Genomics Viewer). All data can be downloaded from http://ftp.pride.ebi.ac.uk/pride/data/proteogenomics/latest/archive/, for downstream analysis.

Moving data into Expression Atlas: re-analysis of quantitative datasets

At the time of writing, 15 quantitative proteomics datasets have already been integrated into the Expression Atlas, an EMBL-EBI added value database that provides information about gene and protein expression in different species and contexts (11). All PRIDE integrated proteomics datasets were manually curated and re-analyzed using a MaxQuant based pipeline. By September 2018, five mouse datasets (showcasing a complete proteome, e.g. https://www.ebi.ac.uk/gxa/experiments/E-PROT-16/Results), six datasets coming from cancer cell lines (showcasing the integration between proteomics and transcriptomics data) and four datasets coming from clinical tumor samples had already been integrated in Expression Atlas. From Expression Atlas, in the near future, we plan that relevant quantitative proteomics data will be disseminated into the Open Targets platform (37).

D448 Nucleic Acids Research, 2019, Vol. 47, Database issue

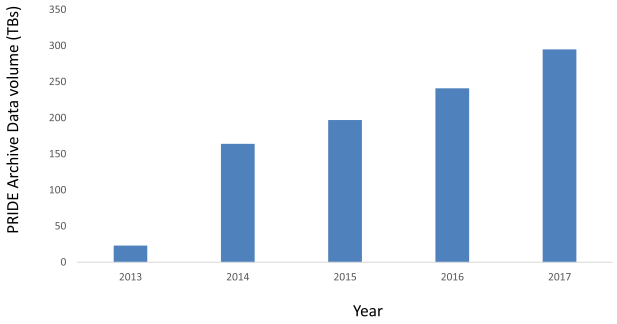


Figure 5. Data volume (in terabytes) downloaded from PRIDE Archive per year.

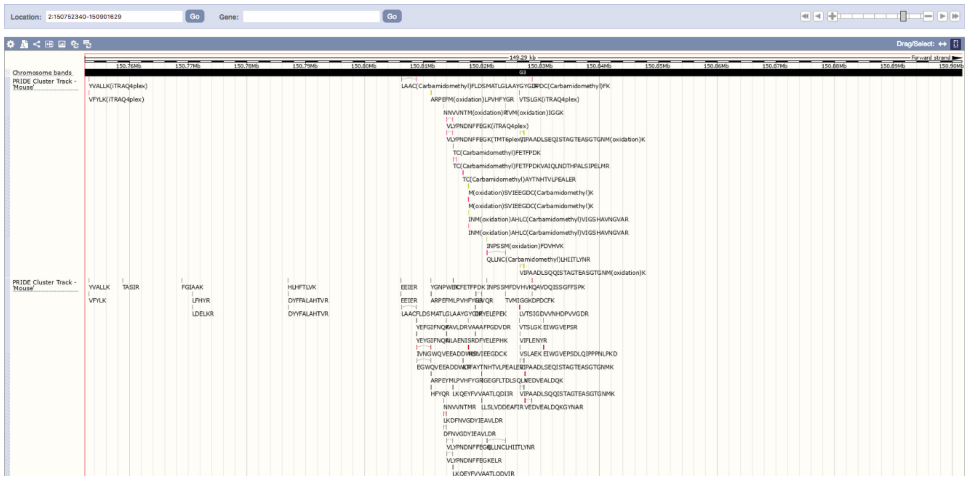


Figure 6. Screenshot of the Ensembl genome browser showing the visualization of peptide evidence as 'TrackHubs' coming from PRIDE Archive and PRIDE Cluster (now PRIDE Peptidome). All peptides shown come from mouse data (GRCm38).

DISCUSSION AND FUTURE PLANS

Thanks, among other efforts, to the success of PRIDE (and of the PX consortium as a whole), the proteomics community is now widely embracing open data policies, an opposite scenario to the situation just a few years ago. In parallel, public proteomics data are being increasingly re-used, with multiple applications. We next outline some of the main working areas for PRIDE in the near future.

First of all, a key aspect is to improve the annotation of the datasets. The current requirements were set up in 2011 (minor updates in 2013), during the establishment of PX, reflecting the discussions at the time, involving many key stakeholders in the field. The main priority was to make data sharing popular. Once this has been achieved, it is now the right time to 'raise the bar'. At the time of writing, a novel annotation system is under development (Supplementary Note 2). The aim is to improve the capture of the experimental design information and technical metadata (e.g. search parameters and relevant information contained in the raw files) (28,38). The improvement in annotation is also required to facilitate further data re-use for third parties. Another key aspect in making data re-use easier is to bring the analysis tools closer to the data, as datasets keep increasing in size.

We are already working in developing open and reproducible data analysis pipelines of different flavours of proteomics workflows (e.g. DDA, DIA, proteogenomics). The main rationale is to make possible the use of that software in cloud infrastructures (using the EMBL-EBI cloud as the starting point), so that in the future the pipelines can be used by the community in the cloud using software container technologies (39,40). In addition, we aim to increasingly perform internal data re-use (including data reprocessing) and disseminate high-quality proteomics data

Nucleic Acids Research, 2019, Vol. 47, Database issue ***D449***

from PRIDE into the already mentioned added-value resources (Ensembl, UniProt and Expression Atlas), among others. At present, identified proteins in PRIDE 'Complete' datasets are cross-referenced in the corresponding UniProt entries (e.g. https://www.uniprot.org/uniprot/Q12181) and 'TrackHubs' are published for some 'Complete' datasets in Ensembl. We plan to enable a more detailed annotation of UniProt and Ensembl entries using proteomics evidence coming from PRIDE, focusing on PTMs, sequence variants and quantitative expression information.

To support this, integration of re-analyzed datasets and the corresponding results in the PRIDE Archive infrastructure needs to be properly supported. Another highly relevant topic for the coming years is the management of clinical proteomics data, and whether they should be considered as patient identifiable or not. This topic has recently gained more relevance after the introduction of the GDPR (General Data Protection Regulation) guidelines by the European Union and we plan to discuss it further in the context of the ELIXIR activities (https://www.elixir-europe.org/). In this context, it is important to highlight that in 2017, PRIDE was named an ELIXIR core data resource (https://www.elixir-europe.org/platforms/data/core-data-resources), joining those biological databases considered to be essential for the scientific community, highlighting the need to make them sustainable in the long term (41).

To finalize, we invite interested parties in PRIDE related developments to follow the PRIDE Twitter account (@pride_ebi). For regular announcements of all the new publicly available datasets, users can follow the PX Twitter account (@proteomexchange).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank all the members of the PRIDE Scientific Advisory Board during the period 2015 to 2018, namely Ruedi Aebersold, Roz Banks, Jurgen Cox, Pedro Cutillas, Concha Gil, Angus Lamond, Kathryn Lilley, Juri Rappsilber, Hans Vissers and Ioannis Xenarios. We would also like to express our sincere gratitude to Henning Hermjakob. At last, we would like to thank all data submitters and collaborators for their contributions.

FUNDING

Wellcome Trust [WT101477MA, 208391/Z/17/Z]; BBSRC Grants 'PROCESS' [BB/K01997X/1]; 'ProteoGenomics' [BB/L024225/1]; 'Proteomics DIA' [BB/P024599/1]; UK-Japan Partnership Award [BB/N022440/1]; NIH 'Proteomics Standards' Grant [R24 GM127667-01]; EU H2020 Project THOR [654039]; Three ELIXIR Implementation Studies; EMBL Core Funding; de.NBI Project of the German Federal Ministry of Education and Research (BMBF) [FKZ 031 A 534A to G.M., 031 A 535A to J.P., T.S.]; Vienna Science and Technology Fund (WWTF) [LS11-045 to J.G.]; PURE (Protein research Unit Ruhr within Europe), a project of North Rhine-Westphalia, a federal state of Germany (to M.E.); European Union's Horizon 2020 Research and Innovation Program [686547 to J.C.]; FP7 [GA ERC-2012-SyG_318987-ToPAG to S.T.]. Funding for open access charge: Wellcome Trust.
Conflict of interest statement. None declared.

REFERENCES

1. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
2. Vizcaino,J.A., Csordas,A., del-Toro,N., Dianes,J.A., Griss,J., Lavidas,I., Mayer,G., Perez-Riverol,Y., Reisinger,F., Ternent,T. *et al.* (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.*, **44**, D447–D456.
3. Deutsch,E.W., Csordas,A., Sun,Z., Jarnuczak,A., Perez-Riverol,Y., Ternent,T., Campbell,D.S., Bernal-Llinares,M., Okuda,S., Kawano,S. *et al.* (2017) The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.*, **45**, D1100–D1106.
4. Deutsch,E.W., Lam,H. and Aebersold,R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.*, **9**, 429–434.
5. Farrah,T., Deutsch,E.W., Kreisberg,R., Sun,Z., Campbell,D.S., Mendoza,L., Kusebauch,U., Brusniak,M.Y., Huttenhain,R., Schiess,R. *et al.* (2012) PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics*, **12**, 1170–1175.
6. Okuda,S., Watanabe,Y., Moriya,Y., Kawano,S., Yamamoto,T., Matsumoto,M., Takami,T., Kobayashi,D., Araki,N., Yoshizawa,A.C. *et al.* (2017) jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.*, **45**, D1107–D1111.
7. Sharma,V., Eckels,J., Schilling,B., Ludwig,C., Jaffe,J.D., MacCoss,M.J. and MacLean,B. (2018) Panorama Public: A public repository for quantitative data sets processed in skyline. *Mol. Cell. Proteomics*, **17**, 1239–1244.
8. Wang,R., Perez-Riverol,Y., Hermjakob,H. and Vizcaino,J.A. (2015) Open source libraries and frameworks for biological data visualisation: a guide for developers. *Proteomics*, **15**, 1356–1374.
9. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Giron,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
10. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, D158–D169.
11. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Fullgrabe,A., Fuentes,A.M., George,N. *et al.* (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.
12. Perez-Riverol,Y., Xu,Q.W., Wang,R., Uszkoreit,J., Griss,J., Sanchez,A., Reisinger,F., Csordas,A., Ternent,T., Del-Toro,N. *et al.* (2016) PRIDE inspector toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. *Mol. Cell. Proteomics*, **15**, 305–317.
13. Reisinger,F., del-Toro,N., Ternent,T., Hermjakob,H. and Vizcaíno,J.A. (2015) Introducing the PRIDE Archive RESTful web services. *Nucleic Acids Res.*, **43**, W599–W604.
14. Perez-Riverol,Y., Uszkoreit,J., Sanchez,A., Ternent,T., Del Toro,N., Hermjakob,H., Vizcaino,J.A. and Wang,R. (2015) ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. *Bioinformatics*, **31**, 2903–2905.
15. Uszkoreit,J., Maerkens,A., Perez-Riverol,Y., Meyer,H.E., Marcus,K., Stephan,C., Kohlbacher,O. and Eisenacher,M. (2015) PIA: an intuitive protein inference engine with a web-based user interface. *J. Proteome Res.*, **14**, 2988–2997.
16. Audain,E., Uszkoreit,J., Sachsenberg,T., Pfeuffer,J., Liang,X., Hermjakob,H., Sanchez,A., Eisenacher,M., Reinert,K., Tabb,D.L. *et al.* (2017) In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J. Proteomics*, **150**, 170–182.
17. Deutsch,E.W., Orchard,S., Binz,P.A., Bittremieux,W., Eisenacher,M., Hermjakob,H., Kawano,S., Lam,H., Mayer,G., Menschaert,G. *et al.* (2017) Proteomics standards initiative: Fifteen years of progress and future work. *J. Proteome Res.*, **16**, 4288–4298.

Chapter 3

Discussion and Outlook

Machine learning and deep learning methods are now emerging as an essential tool in analyzing life science datasets (e.g. medical imaging, genomics, and proteomics). In this thesis, we aimed to predict MS/MS spectra intensities given the peptide sequence. We developed two regression models to predict MS/MS spectra: Deepmass:Prism and wiNNer. With DeepMass:Prism, which is a bidirectional LSTM model, the predicted intensities were as accurate as of the limits of technical reproducibility. Deep learning methods are usually computationally expensive, time-consuming, and require huge datasets as input. For these reasons, we needed a simpler machine learning algorithm, which can be trained easily and efficiently with smaller datasets. A sliding window-based method using conventional neural networks was developed named wiNNer. It has slightly inferior predictive performance compared to DeepMass:Prism, but is computationally inexpensive to train.

In the past deep learning models were known as black-boxes but now can be interpreted using tools like integrated gradient. The results in *article 1* show how each amino acid residue contributes to peptide fragment intensities. Predicted MS/MS spectra can help DDA and DIA computational workflows to improve peptide identification rates and be independent of spectral libraries. For both DDA and DIA applications, the integration of intensity prediction into the MaxQuant environment is currently in progress. Both DeepMass:Prism and wiNNer were implemented using libraries (Keras and Tensorflow) written in python. To deploy these models in the MaxQuant environment (written in C#), it was necessary to use the machine learning libraries compatible with C#. For this purpose, we tried several libraries such as KerasSharp, TensorflowSharp,

Sharplearning, and CNTK. The predictive performance and training speed of wiNNer implemented using Sharplearning is similar to the one implemented in python. Unidirectional LSTM layer in CNTK currently works in C#. This can already be used for example to predict retention time. In the future, the bidirectional LSTM layer implementation can also train the intensity prediction model in C#. Currently, we use the tab separated file generated from batch prediction mode of DeepMass:Prism to create the *in silico* spectral library. Online training of wiNNer in the MaxQuant, will make it possible to take predicted intensities for each peptide of the current dataset within the MaxQuant run and include it in Andromeda score calculation. MaxQuant software now provides a MaxDIA platform to analyzeDIA proteomics data (manuscript is submitted). MaxDIA achieves cutting-edge results with both spectral library and *in silico*predicted spectral libraries. In *article 2*, we showed that the wiNNer model trained on tryptic peptides can be easily adapted for non-traditional proteomics datasets like paleoproteomics samples that have specific characteristics, such as unknown post-translational modifications. The successful applications show that mass spectrometric data analysis will benefit a lot from the predicted spectra in the future.

Machine learning-based predictors such as DISOPRED, PSIPRED are constantly being updated using the latest algorithms and can be used to extract features that can explain phenomena like protein aggregation in neurodegenerative disorders such as Alzheimer's and Huntington's disease. PRIDE is one of the largest repositories for mass spectrometry datasets and it is constantly improving its resources and tools. The mzTab output format can also be adapted for both DDA and DIA data.

In conclusion, deep learning algorithms have immense potential in the future to understand mass spectrometry, protein, and proteomics datasets.

Acronyms

aa amino acid

BP Back propagation

CID collision-induced dissociation

DDA data-dependent acquisition

DIA data-independent acquisition

ECD electron-capture dissociation

ESI electrospray ionization

ETD electron-transfer dissociation

FDR false discovery rate

GC granular component

GPU Graphical processing units

GRU gated recurrent unit

HCD higher-energy collision dissociation

HMM hidden Markov model

HPLC high performance liquid chromatography

HS heat shock

LC liquid chromatography

LSTM long short term memory

MALDI matrix-assisted laser desorption/ionization

MS mass spectrometry

MS^1 full scan

MS^2 fragment scan

PCC Pearson correlation coefficient

PEP posterior error probability

PSM peptide spectrum match

PTMs post-translational modifications

RF random forest

RNN recurrent neural networks

RP reversed-phase

SCX strong cation exchange

SGD Stochastic gradient descent

SILAC stable isotope labeling by amino acids in cell culture

SVM support vector machines

Bibliography

[1] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, aug 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3547.

[2] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 114(28720701):8247–8252, August 2017. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5547637/.

[3] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46:2699, Mar 2018.

[4] Yasset Perez-Riverol, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J. Kundu, Avinash Inuganti, Johannes Griss, Gerhard Mayer, Martin Eisenacher, Enrique Pérez, Julian Uszkoreit, Julianus Pfeuffer, Timo Sachsenberg, Şule Yılmaz, **Shivani Tiwary**, Jürgen Cox, Enrique Audain, Mathias Walzer, Andrew F. Jarnuczak, Tobias Ternent, Alvis Brazma, and Juan Antonio Vizcaíno. The pride database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research*, 47(D1): D442–D450, nov 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1106.

[5] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, nov 2008. ISSN 1087-0156. doi: 10.1038/nbt.1511.

[6] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones. The disopred

server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–2139, mar 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth195.

[7] Alain Coletta, John W. Pinney, David Y. Weiss Solís, James Marsh, Steve R. Pettifer, and Teresa K. Attwood. Low-complexity regions within protein sequences have position-dependent roles. *BMC Systems Biology*, 4(1), apr 2010. ISSN 1752-0509. doi: 10.1186/1752-0509-4-43.

[8] Dirk A. Wolters, Michael P. Washburn, and John R. Yates. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, 73(23):5683–5690, dec 2001. ISSN 0003-2700. doi: 10.1021/ac010617e.

[9] Luca Fornelli, Kenneth R Durbin, Ryan T Fellers, Bryan P Early, Joseph B Greer, Richard D LeDuc, Philip D Compton, and Neil L Kelleher. Advancing top-down analysis of the human proteome using a benchtop quadrupole-orbitrap mass spectrometer. *Journal of proteome research*, 16(2):609–618, 2017.

[10] Timothy K Toby, Luca Fornelli, and Neil L Kelleher. Progress in top-down proteomics and the analysis of proteoforms. *Annual review of analytical chemistry*, 9: 499–519, 2016.

[11] Brian T. Chait. Mass spectrometry: Bottom-up or top-down? *Science*, 314 (5796):65–66, 2006. ISSN 0036-8075. doi: 10.1126/science.1133987. URL https://science.sciencemag.org/content/314/5796/65.

[12] Jungkap Park, Paul D. Piehowski, Christopher Wilkins, Mowei Zhou, Joshua Mendoza, Grant M. Fujimoto, Bryson C. Gibbons, Jared B. Shaw, Yufeng Shen, Anil K. Shukla, Ronald J. Moore, Tao Liu, Vladislav A. Petyuk, Nikola Tolić, Ljiljana Paša-Tolić, Richard D. Smith, Samuel H. Payne, and Sangtae Kim. Informed-proteomics: open-source software package for top-down proteomics. *Nature Methods*, 14(9):909–914, aug 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4388.

[13] Stefka Tyanova, Tikira Temu, and Juergen Cox. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, 11 (12):2301–2319, oct 2016. ISSN 1754-2189. doi: 10.1038/nprot.2016.136.

[14] David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using

mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, dec 1999. ISSN 0173-0835. doi: 10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2.

[15] 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. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, nov 1994. ISSN 1044-0305. doi: 10.1016/1044-0305(94)80016-2.

[16] Robertson Craig and Ronald C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*, 20:1466–7, Jun 2004.

[17] Kirti Sharma, Rochelle C. J. D'Souza, Stefka Tyanova, Christoph Schaab, Jacek R. Wiśniewski, Jürgen Cox, and Matthias Mann. Ultradeep human phosphoproteome reveals a distinct regulatory nature of tyr and ser/thr-based signaling. *Cell Reports*, 8(5):1583–1594, sep 2014. ISSN 2211-1247. doi: 10.1016/j.celrep.2014.07.036.

[18] Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. The perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods*, 13(9): 731–740, 2016.

[19] Mihaela Pertea and Steven L. Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11(5):206, May 2010. ISSN 1474-760X. URL https://doi.org/10.1186/gb-2010-11-5-206.

[20] George A. Khoury, Richard C. Baliban, and Christodoulos A. Floudas. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports*, 1(1), sep 2011. ISSN 2045-2322. doi: 10.1038/srep00090.

[21] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, nov 2008. ISSN 1061-4036. doi: 10.1038/ng.259.

[22] Nagarjuna Nagaraj, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, and Matthias Mann. Deep proteome and

transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7 (1):548, 2011.

[23] Martin Beck, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, and Ruedi Aebersold. The quantitative proteome of a human cell line. *Molecular systems biology*, 7(1):549, 2011.

[24] Michael. Karas and Franz. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20): 2299–2301, oct 1988. ISSN 0003-2700. doi: 10.1021/ac00171a028.

[25] J. Fenn, M. Mann, C. Meng, S. Wong, and C. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, oct 1989. ISSN 0036-8075. doi: 10.1126/science.2675315.

[26] Tamar Geiger, Anja Wehner, Christoph Schaab, Juergen Cox, and Matthias Mann. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & cellular proteomics : MCP*, 11:M111.014050, Mar 2012.

[27] Pavel Sinitcyn, Jan Daniel Rudolph, and Jürgen Cox. Computational methods for understanding mass spectrometry–based shotgun proteomics data. *Annual Review of Biomedical Data Science*, 1(1):207–234, 2018. doi: 10.1146/annurev-biodatasci-080917-013516. URL https://doi.org/10.1146/annurev-biodatasci-080917-013516.

[28] Jacek R Wiśniewski, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. Universal sample preparation method for proteome analysis. *Nature methods*, 6(5):359–362, 2009.

[29] Jesper V Olsen, Shao-En Ong, and Matthias Mann. Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*, 3(6): 608–614, 2004.

[30] Timo Glatter, Christina Ludwig, Erik Ahrne, Ruedi Aebersold, Albert JR Heck, and Alexander Schmidt. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tan-

dem lys-c/trypsin proteolysis over trypsin digestion. *Journal of proteome research*, 11(11):5145–5156, 2012.

[31] Suman S Thakur, Tamar Geiger, Bhaswati Chatterjee, Peter Bandilla, Florian Fröhlich, Juergen Cox, and Matthias Mann. Deep and highly sensitive proteome coverage by lc-ms/ms without prefractionation. *Molecular & cellular proteomics*, 10(8), 2011.

[32] J. Alex Taylor and Richard S. Johnson. Sequence database searches viade novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9):1067–1075, jun 1997. ISSN 0951-4198. doi: 10.1002/(sici)1097-0231(19970615)11:9<1067::aid-rcm953>3.0.co;2-l.

[33] Annette Michalski, Juergen Cox, and Matthias 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. *Journal of proteome research*, 10(4): 1785–1793, 2011.

[34] Lloyd R. Snyder, Joseph Jack Kirkland, and John W. Dolan. *Introduction to modern liquid chromatography*. Wiley, Hoboken, NJ, 3rd. ed. edition, 2010. ISBN 9780470167540. URL https://www.ebook.de/de/product/9395223/snyder_dolan_kirkland_john_w_dolan_liquid_chromatography_3e.html. Literaturangaben.

[35] James W Jorgenson. Capillary liquid chromatography at ultrahigh pressures. *Annual review of analytical chemistry*, 3:129–150, 2010.

[36] Richard D Smith, Yufeng Shen, and Keqi Tang. Ultrasensitive and quantitative analyses from combined separations- mass spectrometry for the characterization of proteomes. *Accounts of chemical research*, 37(4):269–278, 2004.

[37] Geoffrey Ingram Taylor. Disintegration of water drops in an electric field. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 280(1382):383–397, 1964.

[38] Matthias Mann and Matthias Wilm. Electrospray mass spectrometry for protein characterization. *Trends in biochemical sciences*, 20(6):219–224, 1995.

[39] Alan G Marshall, Christopher L Hendrickson, and George S Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass spectrometry reviews*, 17(1):1–35, 1998.

[40] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. The orbitrap: a new mass spectrometer. *Journal of mass spectrometry*, 40(4):430–443, 2005.

[41] Jae C Schwartz, Michael W Senko, and John E.P Syka. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry*, 13(6):659 – 669, 2002. ISSN 1044-0305. doi: https://doi.org/10.1016/S1044-0305(02)00384-7. URL http://www.sciencedirect.com/science/article/pii/S1044030502003847.

[42] Alexander Makarov and Eduard Denisov. Dynamics of ions of intact proteins in the orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry*, 20:1486–95, Aug 2009.

[43] Roman A. Zubarev and Alexander Makarov. Orbitrap mass spectrometry. *Analytical Chemistry*, 85(11):5288–5296, 2013. doi: 10.1021/ac4001223. URL https://doi.org/10.1021/ac4001223. PMID: 23590404.

[44] Jesper V Olsen, Jae C Schwartz, Jens Griep-Raming, Michael L Nielsen, Eugen Damoc, Eduard Denisov, Oliver Lange, Philip Remes, Dennis Taylor, Maurizio Splendore, Eloy R Wouters, Michael Senko, Alexander Makarov, Matthias Mann, and Stevan Horning. A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. *Molecular & cellular proteomics : MCP*, 8:2759–2769, December 2009. ISSN 1535-9484. doi: 10.1074/mcp.M900375-MCP200.

[45] Matthias Mann and Neil L. Kelleher. Precision proteomics: The case for high resolution and high mass accuracy. *Proc Natl Acad Sci USA*, 105(47):18132, November 2008. URL http://www.pnas.org/content/105/47/18132.abstract.

[46] Vicki H Wysocki, George Tsaprailis, Lori L Smith, and Linda A Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35(12):1399–1406, 2000.

[47] Graeme C. McAlister, Douglas H. Phanstiel, Justin Brumbaugh, Michael S. Westphall, and Joshua J. Coon. Higher-energy collision-activated dissociation without a dedicated collision cell. *Mol Cell Proteomics*, 10(5):O111.009456, May 2011. URL http://www.mcponline.org/content/10/5/O111.009456.abstract.

[48] Annette Michalski, Nadin Neuhauser, Jürgen Cox, and Matthias Mann. A systematic investigation into the nature of tryptic hcd spectra. *Journal of Proteome Research*, 11(11):5479–5491, oct 2012. ISSN 1535-3893. doi: 10.1021/pr3007045.

[49] J. Mitchell Wells and Scott A. McLuckey. Collision-induced dissociation (cid) of peptides and proteins, 2005. ISSN 0076-6879.

[50] Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5(9):699–711, sep 2004. ISSN 1471-0072. doi: 10.1038/nrm1468.

[51] Jesper V. Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, aug 2007. ISSN 1548-7091. doi: 10.1038/nmeth1060.

[52] P Roepstorff and J Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry*, 11:601, November 1984. ISSN 0306-042X. doi: 10.1002/bms.1200111109.

[53] K Biemann. Contributions of mass spectrometry to peptide and protein structure. *Biomedical & environmental mass spectrometry*, 16:99–111, October 1988. ISSN 0887-6134. doi: 10.1002/bms.1200160119.

[54] R A Zubarev, D M Horn, E K Fridriksson, N L Kelleher, N A Kruger, M A Lewis, B K Carpenter, and F W McLafferty. Electron capture dissociation for structural characterization of multiply charged protein cations. *Analytical chemistry*, 72:563–573, February 2000. ISSN 0003-2700. doi: 10.1021/ac990811p.

[55] John E. P. Syka, Joshua J. Coon, Melanie J. Schroeder, Jeffrey Shabanowitz, and Donald F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*, 101(26):9528, June 2004. URL http://www.pnas.org/content/101/26/9528.abstract.

[56] George Rosenberger, Isabell Bludau, Uwe Schmitt, Moritz Heusel, Christie L. Hunter, Yansheng Liu, Michael J. MacCoss, Brendan X. MacLean, Alexey I. Nesvizhskii, Patrick G. A. Pedrioli, Lukas Reiter, Hannes L. Röst, Stephen Tate, Ying S. Ting, Ben C. Collins, and Ruedi Aebersold. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nature Methods*, 14(9):921–927, aug 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4398.

[57] Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, and Alexey I. Nesvizhskii. Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, 12(3):258–264, jan 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3255.

[58] Jemma X. Wu, Xiaomin Song, Dana Pascovici, Thiri Zaw, Natasha Care, Christoph Krisp, and Mark P. Molloy. Swath mass spectrometry performance using extended peptide ms/ms assay libraries. *Molecular & Cellular Proteomics*, 15(7):2501–2514, may 2016. ISSN 1535-9476. doi: 10.1074/mcp.m115.055558.

[59] Allison Doerr. Dia mass spectrometry. *Nature Methods*, 12(1):35–35, dec 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3234.

[60] Lewis Y. Geer, Sanford P. Markey, Jeffrey A. Kowalak, Lukas Wagner, Ming Xu, Dawn M. Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H. Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3:958–64, Sep-Oct 2004.

[61] Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. Andromeda: A peptide search engine integrated into the maxquant environment. *Journal of Proteome Research*, 10(4):1794–1805, 2011. doi: 10.1021/pr101065j. URL https://doi.org/10.1021/pr101065j. PMID: 21254760.

[62] Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, and Fiona Cunningham. Ensembl variation resources. *Database*,

2018, 11 2018. ISSN 1758-0463. doi: 10.1093/database/bay119. URL https://doi.org/10.1093/database/bay119. bay119.

[63] Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, feb 2007. ISSN 1548-7091. doi: 10.1038/nmeth1019.

[64] Jürgen Cox, Ivan Matic, Maximiliane Hilger, Nagarjuna Nagaraj, Matthias Selbach, Jesper V. Olsen, and Matthias Mann. A practical guide to the maxquant computational platform for silac-based quantitative proteomics. *Nature protocols*, 4:698–705, 2009.

[65] Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Analytical Chemistry*, 74(20):5383–5392, oct 2002. ISSN 0003-2700. doi: 10.1021/ac025747h.

[66] Hyungwon Choi and Alexey I. Nesvizhskii. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(1):254–265, jan 2008. ISSN 1535-3893. doi: 10.1021/pr070542g.

[67] Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, oct 2007. ISSN 1548-7091. doi: 10.1038/nmeth1113.

[68] Sven Degroeve and Lennart Martens. Ms2pip: a tool for ms/ms peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, sep 2013. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt544.

[69] Sven Degroeve, Davy Maddelein, and Lennart Martens. Ms2pip prediction server: compute and visualize ms2peak intensity predictions for cid and hcd fragmentation. *Nucleic Acids Research*, 43(W1):W326–W330, may 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv542.

[70] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for pep-

tidede novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342, 2003. ISSN 0951-4198. doi: 10.1002/rcm.1196.

[71] Bin Ma and Richard Johnson. De novo sequencing and homology searching. *Molecular & cellular proteomics : MCP*, 11:O111.014902, Feb 2012.

[72] Xiaoyu Yang, Vijay Dondeti, Rebecca Dezube, Dawn M Maynard, Lewis Y Geer, Jonathan Epstein, Xiongfong Chen, Sanford P Markey, and Jeffrey A Kowalak. Dbparser: web-based software for shotgun proteomic data analyses. *Journal of proteome research*, 3:1002–1008, 2004. ISSN 1535-3893. doi: 10.1021/pr049920x.

[73] Ze-Qiang Ma, Surendra Dasari, Matthew C Chambers, Michael D Litton, Scott M Sobecki, Lisa J Zimmerman, Patrick J Halvey, Birgit Schilling, Penelope M Drake, Bradford W Gibson, and David L Tabb. Idpicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of proteome research*, 8:3872–3881, August 2009. ISSN 1535-3893. doi: 10.1021/pr900360j.

[74] Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 75:4646–4658, September 2003. ISSN 0003-2700. doi: 10.1021/ac0341261.

[75] Jacek R Wiśniewski, Marco Y Hein, Juergen Cox, and Matthias Mann. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Molecular & cellular proteomics*, 13(12):3497–3506, 2014.

[76] Shao-En Ong and Matthias Mann. *Stable Isotope Labeling by Amino Acids in Cell Culture for Quantitative Proteomics*, pages 37–52. Humana Press, Totowa, NJ, 2007. ISBN 978-1-59745-255-7. doi: 10.1007/978-1-59745-255-7_3. URL https://doi.org/10.1007/978-1-59745-255-7_3.

[77] Haining Zhu, Songqin Pan, Sheng Gu, E Morton Bradbury, and Xian Chen. Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Communications in Mass Spectrometry*, 16(22):2115–2123, 2002.

[78] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags:

a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical chemistry*, 75(8):1895–1904, 2003.

[79] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389:1017–1031, October 2007. ISSN 1618-2642. doi: 10.1007/s00216-007-1486-6.

[80] Marcus Bantscheff, Simone Lemeer, Mikhail M Savitski, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry*, 404:939–965, September 2012. ISSN 1618-2650. doi: 10.1007/s00216-012-6203-4.

[81] Hongbin Liu, Rovshan G Sadygov, and John R Yates. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry*, 76(14):4193–4201, 2004.

[82] Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular & cellular proteomics*, 13(9):2513–2526, 2014.

[83] Vladimir N. Vapnik. The nature of statistical learning theory, 1995.

[84] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 0028-0836. doi: 10.1038/nature16961.

[85] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[86] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[87] Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[88] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997. URL http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf.

[89] Stefka Tyanova, Reidar Albrechtsen, Pauliina Kronqvist, Juergen Cox, Matthias Mann, and Tamar Geiger. Proteomic maps of breast cancer subtypes. *Nature communications*, 7(1):1–11, 2016.

[90] Daniel N Itzhak, Stefka Tyanova, Jürgen Cox, and Georg Hh Borner. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife*, 5, June 2016. ISSN 2050-084X. doi: 10.7554/eLife.16950.

[91] Daniel N Itzhak, Colin Davies, Stefka Tyanova, Archana Mishra, James Williamson, Robin Antrobus, Jürgen Cox, Michael P Weekes, and Georg HH Borner. A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell reports*, 20(11):2706–2718, 2017.

[92] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. URL https://doi.org/10.1023/A:1010933404324.

[93] Joshua E Elias, Francis D Gibbons, Oliver D King, Frederick P Roth, and Steven P Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature biotechnology*, 22(2):214–219, 2004.

[94] Warren S. McCulloch and Walter Pitts. *A Logical Calculus of the Ideas Immanent in Nervous Activity*, page 15–27. MIT Press, Cambridge, MA, USA, 1988. ISBN 0262010976.

[95] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[96] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v15/glorot11a.html.

[97] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[98] T. Tieleman and G Hinton. Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. Technical report, 2012.

[99] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[100] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[101] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1487–1495, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098043. URL https://doi.org/10.1145/3097983.3098043.

[102] François Chollet et al. Keras. https://keras.io, 2015.

[103] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA, November 2016. USENIX Association. ISBN 978-1-931971-33-1. URL https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

[104] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, nov 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

[105] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[106] H Nielsen, J Engelbrecht, S Brunak, and G von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein engineering*, 10:1–6, January 1997. ISSN 0269-2139. doi: 10.1093/protein/10.1.1.

[107] Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

[108] J C Wootton and S Federhen. Analysis of compositionally biased regions in sequence databases. *Methods in enzymology*, 266:554–571, 1996. ISSN 0076-6879. doi: 10.1016/s0076-6879(96)66035-2.

[109] Henrik Nielsen and Anders Krogh. Prediction of signal peptides and signal anchors by a hidden markov model. In *Ismb*, volume 6, pages 122–130, 1998.

[110] José Juan Almagro Armenteros, Konstantinos D Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, 37:420–423, April 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0036-z.

[111] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, nov 1999. ISSN 1367-4803. doi: 10.1093/bioinformatics/15.11.937.

[112] Ivica Letunic and Peer Bork. 20 years of the smart protein domain annotation resource. *Nucleic acids research*, 46:D493–D496, January 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx922.

[113] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The pfam protein families database in 2019. *Nucleic acids research*, 47:D427–D432, January 2019. ISSN 1362-4962. doi: 10.1093/nar/gky995.

[114] Pawel Smialowski, Thorsten Schmidt, Jürgen Cox, Andreas Kirschner, and Dmitrij Frishman. Will my protein crystallize? a sequence-based predictor. *Proteins: Structure, Function, and Bioinformatics*, 62(2):343–355, 2006.

[115] Robert Boyd and Árpád Somogyi. The mobile proton hypothesis in fragmentation of protonated peptides: A perspective. *Journal of the American Society for Mass Spectrometry*, 21(8):1275–1278, aug 2010. ISSN 1044-0305. doi: 10.1016/j.jasms.2010.04.017.

[116] R. A. N. D. Y. J. ARNOLD, N. A. R. M. A. D. A. JAYASANKAR, D. I. V. Y. A. AGGARWAL, H. A. I. X. U. TANG, and P. R. E. D. R. A. G. RADIVOJAC. A machine learning approach to predicting peptide fragmentation spectra, dec 2005.

[117] Sujun Li, Randy J. Arnold, Haixu Tang, and Predrag Radivojac. On the accuracy and limits of peptide fragmentation spectrum prediction. *Analytical Chemistry*, 83 (3):790–796, feb 2011. ISSN 0003-2700. doi: 10.1021/ac102272r.

[118] Zhongqi Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry*, 76(14):3908–3922, 2004.

[119] Zhongqi Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Analytical chemistry*, 77(19):6364–6373, 2005.

[120] Xie-Xuan Zhou, Wen-Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si-Min He, and Zhifei Zhang. pdeep: Predicting ms/ms spectra of peptides with deep learning. *Analytical Chemistry*, 89(23):12690–12697, nov 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.7b02566.

[121] Christian D. Kelstrup, Dorte B. Bekker-Jensen, Tabiwang N. Arrey, Alexander Hogrebe, Alexander Harder, and Jesper V. Olsen. Performance evaluation of the

q exactive hf-x for shotgun proteomics. *Journal of Proteome Research*, 17(1):727–738, dec 2018. ISSN 1535-3893. doi: 10.1021/acs.jproteome.7b00602.

[122] **Tiwary, Shivani**, Roie Levy, Petra Gutenbrunner, Favio Salinas Soto, Krishnan K. Palaniappan, Laura Deming, Marc Berndl, Arthur Brant, Peter Cimermancic, and Jürgen Cox. High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, 16(6):519–525, June 2019. ISSN 1548-7105. URL https://doi.org/10.1038/s41592-019-0427-6.

[123] Frido Welker, Jazmín Ramos-Madrigal, Petra Gutenbrunner, Meaghan Mackie, **Tiwary, Shivani**, Rosa Rakownikow Jersie-Christensen, Cristina Chiva, Marc R. Dickinson, Martin Kuhlwilm, Marc de Manuel, Pere Gelabert, María Martinón-Torres, Ann Margvelashvili, Juan Luis Arsuaga, Eudald Carbonell, Tomas Marques-Bonet, Kirsty Penkman, Eduard Sabidó, Jürgen Cox, Jesper V. Olsen, David Lordkipanidze, Fernando Racimo, Carles Lalueza-Fox, José María Bermúdez de Castro, Eske Willerslev, and Enrico Cappellini. The dental proteome of homo antecessor. *Nature*, 580(7802):235–238, April 2020. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-020-2153-8.

[124] F. Frottin, F. Schueder, **Tiwary, S.**, R. Gupta, R. Körner, T. Schlichthaerle, J. Cox, R. Jungmann, F. U. Hartl, and M. S. Hipp. The nucleolus functions as a phase-separated protein quality control compartment. *Science*, 365(6451):342–347, 2019. ISSN 0036-8075. doi: 10.1126/science.aaw9157. URL https://science.sciencemag.org/content/365/6451/342.

[125] Johannes Griss, Andrew R. Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G. Thallinger, Reza M. Salek, Christoph Steinbeck, Nadin Neuhauser, Jürgen Cox, Steffen Neumann, Jun Fan, Florian Reisinger, Qing-Wei Xu, Noemi del Toro, Yasset Pérez-Riverol, Fawaz Ghali, Nuno Bandeira, Ioannis Xenarios, Oliver Kohlbacher, Juan Antonio Vizcaíno, and Henning Hermjakob. The mztab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics*, 13(10):2765–2775, jun 2014. ISSN 1535-9476. doi: 10.1074/mcp.o113.036681.

Acknowledgments

I would like to express my gratitude to the numerous people who made this thesis possible. I want to thank Prof. Dr. Thomas Carrel, who was so kind to agree to be my Doctorvater and serving on my thesis examination committee, Your help was invaluable.

A very special thanks to Dr. Jürgen Cox for giving me the opportunity to work with him on numerous challenging projects. These projects not only helped me to learn every minor detail about the subject but also allowed me to use new technologies in both the proteomics and machine learning field. I would like to thank him for, constant support and advice. It was a pleasure to learn the essentials of computational mass spectrometry from him. Thank you for providing a nice and interactive working environment. Thank you for giving me the teaching opportunities at summer schools.

Further, I would also like to express my gratitude towards my project collaborators Dr. Peter Cimermancic, Dr. Yasset Perez-Riverol, Dr. F. Frottin, Dr. Débora Trentini, and Dr. Kirti Sharma for the exciting projects, assistance, and fruitful discussions. These collaborations helped me shape my career decisions and future scientific interests.

Special thanks to my project partners Favio, Petra, Daniela, and Sule. Working together with you guys was the best learning experience for me. Special thanks to Pelagia for listening to me nonstop, be it science or life problems, and for all the fun and philosophical talks we had and are still having. Special thanks to Pavel and Daniela for their constant support, jokes, and scientific discussions. Special thanks to Jan for introducing me to the world of fitness and rap music.

The current and past Cox lab members made the past years so much more enjoyable and for all the good times we shared, for stimulating discussions, for their help and support during my thesis, and the nice coffee breaks we shared. Besides the development of my scientific and research skills, I also made friends, whom I will cherish all my life.

I am deeply grateful to Dipali and Gurnoor for insisting me to come back to Germany and apply for Ph.D., and for their constant support and motivation. I would like to thank Prof. Jörg Zimmerman to introduce me to the world of deep learning before it became a buzz word.

I would like to extend my gratitude towards my co-workers (Dolly, Manoj, Kalidoss, Rakesh, Poonam, Ravi, special thanks to late Sudha and our supervisor Dr.Ravi Sirdeshmukh) in CCMB, India, with whom I learned the basics of computational proteomics. It was their constant motivation that I decided to do higher studies abroad.

I am deeply grateful to my childhood friends, connected via WhatsApp, for always being there on my side, uplifting my spirits, and for non-stop entertaining and meaningful discussions. I am so thankful to Julie, Alex, Aniketa, Anisha, and Simone for supporting me all the way and making life fun outside the lab.

Last but not the least, I would like to express my heartfelt thanks to my family for always being supportive and encouraging, throughout my studies. I dedicate this work to my father Dr. R.K.Tiwary. Thank you for introducing me to the world of science and research, and motivating me to pursue higher education. I am especially thankful to my mother Sarita Tiwary, for supporting me in all of my decisions and inspiring me to follow my dreams. Thank you both for always believing in me. I am so privileged to be your daughter. Special thanks to my elder brother Sanjay Tiwary, his wife Nutan Tiwary, and my lovely nephew for bringing joy into our lives. Special thanks to my younger brother Shishir Tiwary for his constant support. I love you guys.