Chapter

Pattern Recognition for Mass-Spectrometry-Based Proteomics

Patrick Bangert, Nandha Kumar Balasubramaniam, Carol E. Parker and Christoph H. Borchers

Abstract

Multiomic analysis comprises genomics, proteomics, and metabolomics leads to meaningful insights but necessitates sifting through voluminous amounts of complex data. Proteomics in particular focuses on the end product of gene expression – i.e., proteins. The mass spectrometric approach has proven to be a workhorse for the qualitative and quantitative study of protein interactions as well as post-translational modifications (PTMs). A key component of mass spectrometry (MS) is spectral data analysis, which is complex and has many challenges as it involves identifying patterns across a multitude of spectra in combination with the meta-data related to the origin of the spectrum. Artificial Intelligence (AI) along with Machine Learning (ML), and Deep Learning (DL) algorithms have gained more attention lately for analyzing the complex spectral data to identify patterns and to create networks of value for biomarker discovery. In this chapter, we discuss the nature of MS proteomic data, the relevant AI methods, and demonstrate their applicability. We also show that AI can successfully identify biomarkers and aid in the diagnosis, prognosis, and treatment of specific diseases.

Keywords: proteomics, mass-spectrometry, artificial intelligence, pattern recognition, diagnosis

1. Introduction

Central Dogma theory highlights the flow of information in most living systems from DNA to RNA to Proteins, in which the four bases – adenine (A), cytosine (C), guanine (G), and thymine (T) in DNA or Uracil (U) in RNA) – encode the 20 amino acids that are the building blocks of proteins. Proteomics is essentially the study of 100 s to 1000s of proteins in a single shot, made possible by the advancements in multiplexing technologies.

Proteins are the functional elements of biology, and the information encoded in DNA or RNA represents a potential protein that may or may not actually be manufactured by the cell. DNA sequencing has enabled the development of the field of proteomics by allowing the prediction of the amino acid sequences of the encoded proteins. The expression of a particular gene leads to the formation of a protein which plays a vital role in running the cellular machinery – from providing structural support to catalysis as enzymes, and even controlling gene expression. These proteins, in turn, form the key components of biological pathways, and may be biomarkers or targets for drugs. They can form complexes and operate through cellular pathways, in which even a single point mutation can lead to a disease state.

Amino acids are the building blocks of peptides, and ultimately proteins, which are then folded into unique three-dimensional shapes. Internal bonding within proteins is essential for stabilizing their structure, and the final folded form is essential for the functional activity. Through the years, innumerable techniques have been used to determine the structure of proteins, but lately, artificial intelligence (AI) is being increasingly adopted to determine the shape taken by proteins [1].

The Human Genome Project (HGP) mapped the genes to the entire protein collection as expressed in the approximately 230 cell types [2]. Based on data available in public databases [3], it is estimated that there are about 20,300–20,500 protein-coding human genes. Legrain, et.al, [4] indicated that approximately 30% of these genes lacked experimental evidence at the protein level.

The general applications of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) to proteomics include annotation and MS spectral analysis. This approach started in 2000, but gained momentum after 2006, when it was used for peptide fragmentation [5, 6] and identification [7]. AI deviates from classical statistical analysis in that it can handle huge volumes of data and makes fewer assumptions.

MS data for proteins can be generated from peptides using either the ‘bottom-up” approach [8], or from an intact protein using the “top-down” approach [9]. In 2013, Swan, et al. suggested the use of machine learning for analyzing, and interpreting large sets of MS data [10]. Two areas of applications highlighted included direct analysis of mass spectral peaks, and proteins identified by sequence database search, which involves relative protein quantification. The outcome from this analysis would facilitate the identification of disease biomarkers and the classification samples as applicable to diagnostics. Challenges include protein prediction, quantification, use of ML, and dealing with small sample sizes.

In the early stages of *in silico* design, researchers generated computer models of protein folding, and created 3D models for understanding drug binding mechanisms. Now, AI is being increasingly used in proteomics, and this has gained increased importance due to the current COVID-19 pandemic. This approach can be used to generate vaccines and to keep pace with the rate at which virus is mutating. Thus, AI is playing a larger role in figuring out the shape of proteins [1] with implications for all aspects of drug research. Complex interactions between proteins in living systems can be deciphered by feeding the right inputs to the algorithms, leading to more accurate results by leveraging AI. ML algorithms can learn the relationship between sequencing data and relating it to the function of the proteins. These algorithms can easily predict mutations. As a result, protein engineering is experiencing rapid growth due to the application of AI and Ml algorithms.

Proteomics Vocabulary	**Meaning**
Amino acid	A biological molecule with the general structure NH_2-HCR_1-COOH
Peptide	A short chain of amino acids, coupled thru the amide bond NH_2-HCR_1-CO-NH-HCR_2-COOH

Proteomics Vocabulary	**Meaning**
Protein	A molecule composed of one or more chains of amino acids. Proteins often function as part of protein complexes (see [11] for examples)
Post-translation modification	A modification of a peptide (e.g. acetylation, ubiquitination, phosphorylation, truncation) that occurs after the peptide or protein is synthesized in the cell
Biomarker	A biological molecule whose concentration is correlated with a disease state or condition.
Mass Spectrometry (MS)	A separation technique that separates charged molecules on the basis of their mass-to-charge ratios (m/z values). It can be a qualitative or quantitative technique. MS is a gas-phase technique that can be interfaced with a separation technique such as gas chromatography or liquid chromatography.
Matrix-assisted Laser Desorption/Ionization (MALDI)	One of the important "soft" ionization techniques used in biological MS. Sample is co-evaporated with a matrix – a compound that assists in the transfer of one or more protons to the analyte. Sample is dried, and is the analyte is ionized with a laser.
Electrospray ionization (ESI)	One of the important "soft" ionization techniques used in biological MS. The sample is ionized as it passes thru a high-voltage needle. Typically, 2–3 protons are added to a peptide, while dozens of protons can be transferred to a protein.
Tandem mass spectrometry (MS/MS)	A MS technique that involves two stages of m/z-based separation. First, a mass spectrum is generated. Then, a selected precursor ion is passed into a region of the instrument where it is fragmented, and then a second mass spectrum is generated.

Uncovering biomarkers for drug discovery has been the driving force behind many of the advances in proteomics [12]. The advanced applications of AI, ML, and DL algorithms, have added a new dimension to the understanding of the causes of neurological diseases such as Alzheimer's and Huntington's diseases, to the design and development of monoclonal antibodies to fight various types of cancer, and to the development of vaccines to combat infectious diseases. These biomarker studies, in turn, have their roots in relating the presence or absence of certain proteins to diseases states [13, 14]. The current trend is movement from a "one-size-fits all" model to personalized/precision medicine approach, in which the treatments are based on the needs of the specific individual or sub-populations [15].

2. Introduction to artificial intelligence

Central to artificial intelligence is the idea of a model. A model is a mathematical representation of something. If the thing being modeled is physical, the model is sometimes called a digital twin. The model can have several functions, e.g. it can forecast the current situation into the future, it can calculate aspects of the situation that are difficult to measure from other aspects that are easy to measure, and it can represent sophisticated patterns and make them available for practical use in automating a variety of processes. Models based on the laws of physics or engineering are known as first-principles models and are computational in nature. Models generally have placeholders for numerical values. These parameters or coefficients must be determined in some way. For physics-based models, they must be experimentally determined. In artificial intelligence, these coefficients are "learned" from data using computational recipes called algorithms.

AI Vocabulary	**Meaning**
Artificial Intelligence (AI)	Umbrella term for many diverse methods that convert a dataset into a model.
Machine Learning (ML)	Subset of AI that deals with numerical data, as opposed to image or language data.
Deep Learning (DL)	Modern term for AI that is largely synonymous with AI emphasizing that it uses models of many "layers," which is a sign of model complexity that became practical after about 2010.
Neural Networks	One of the many methods and model types available in AI and ML.
Training	The process of determining model coefficients from data.
Inference	The process of applying a trained model to a novel data point.
Labels	The human-provided information that augments the dataset with the desired result for each data point in the dataset. Using both the data and the labels, the training process attempts to reconstruct the connection between them, allowing the model to compute the output from the data.

Machine learning (ML) or artificial intelligence (AI) is the name given to a large collection of diverse methods that aim to produce models given enough empirical data. They do not require the use of physical laws or the specification of physical characteristics. They determine the dependency of the variables among each other by using the data, and only the data. That is not to say that there is no more need for a human expert. The human expert is essential but the way the expertise is supplied is very different to the first-principles model – for ML or AI the human domain knowledge is supplied in the form of labels. Labels are human-generated manual annotations to empirical data that identify aspects of relevance. In proteomics that might be the name of a certain peptide whose mass spectrometry signature has been determined and so on.

The subject of ML and AI has three main parts. First, it consists of many prototypical model types that could be applied to the data at hand. These are known by names such as neural networks, decision trees, or k-means clustering; we will not dive deeply into what these mean. Second, each of these types comes with several recipes, called algorithms that tell us how to calculate the model coefficients from a data set. This calculation is also called training the model. After training, the initial prototype has been turned into a model for the specific dataset that we provided. Third, the finished model must be deployed so that it can be used. It is generally far easier and quicker to evaluate a model than to train a model. In fact, this is one of the primary features of ML and AI that make it so attractive: Once trained, the model can be used in real-time. However, it needs to be embedded in the right infrastructure to unfold its potential.

Associated with ML and AI are two essential topics that are at the heart of data science. First, the data must be suitably prepared for learning. Second, the resultant model must be adequately tested, and its performance must be demonstrated using rigorous mathematical means. This pre-processing and post-processing before and after ML or AI is applied to round out the scientific part of a data science project.

While the field of AI began in the 1950's, it has had a rocky evolution over time. It is really only after about 2010 that the most recent push, known as deep learning (DL), had wide-spread practical success. It is around 2015 that several tasks could be handled by AI with a higher accuracy than by humans. This was and remains the most important practical ramification that underlines the automation benefit of AI – if AI can do something about as well as humans can do it, then we can get AI to do it more rapidly and at lower cost.

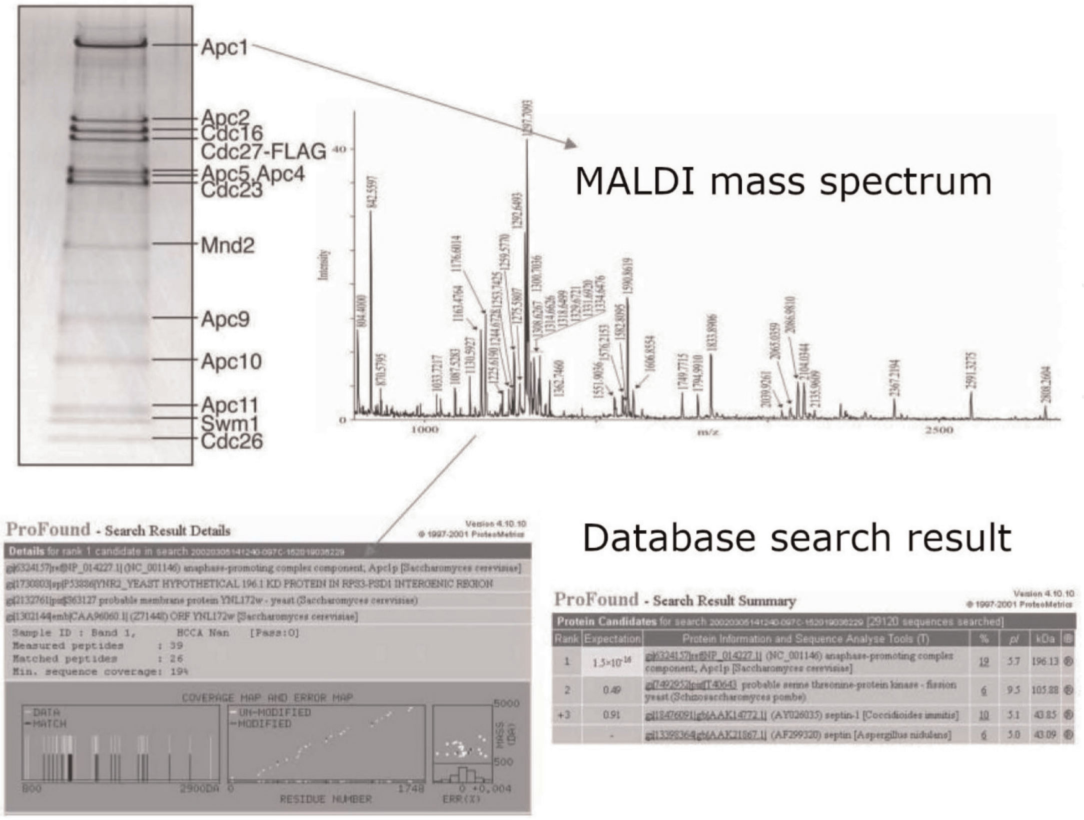
The difference between the three subjects of AI, ML, and DL is very blurred and most people use these terms as synonyms. As such, it is difficult to try to draw a boundary line between them, as most authors will not agree on any one particular demarcation. It is generally accepted, however, that ML deals only with numerical data, so visual and language data are reserved for AI.

3. Protein identification

The proteome is the collection of all proteins present in biofluids, cells, and tissues while proteomics is the study of the proteome. Essentially, in qualitative proteomics, we ask which proteins are present and in quantitative proteomics we ask in what quantities they are present. The first challenges appear when we realize that not all proteins are present in every part of the body in equal measure, for example, in a blood sample. Furthermore, some proteins are present in large quantities while others are present only in trace amounts [16]. In fact, the most common 22 proteins represent over 99% of the total proteome by mass but it is the remaining tens of thousands of proteins that are potentially interesting as biomarkers [17–20]. There is a practical challenge to resolve the interesting part of the proteome in the presence of an overwhelming amount of what, in data analytic terms, is background noise. The dynamic range of proteins in biofluids or cells or tissues presents a challenge to the detection and identification of low-abundance proteins.

3.1 Protein and peptide mass spectra

Working backwards, determining the amino acid sequence, and thereby the mass of a protein, requires breaking it down to peptides. This can be accomplished by chemical or enzymatic methods, with the resulting peptides being electrically charged leading to partial ionization. They are then separated based on their mass-to-charge (m/z) ratios, using electrical and/or magnetic fields [21]. The technology that makes this practicable is called mass spectrometry (MS) which has been the driving force behind proteomics [22, 23]. For the purposes of this paper, we will not go into the physical, biological, or chemical mechanisms underlying this complex procedure. We will simply treat MS as a mechanism that produces a spectrum like shown in **Figure 1**. A peptide mass spectrum can be thought of as a plot of m/z values (on the x-axis) versus intensities (y axis). Intensities are proportional to the amount of material at that m/z ratio, as well as also the sensitivity of that peptide. It is immediately clear from the spectrum in **Figure 1** is that certain m/z values are present in the material and others are absent.



ProFound - Search Result Details

Version 4.10.10
© 1997-2001 ProteoMetrics

Details for rank 1 candidate in search 20020305141240-097C-1520190038229

gi|6324157|ref|NP_014227.1| (NC_001146) anaphase-promoting complex component; Apc1p [Saccharomyces cerevisiae]

gi|1730803|sp|P53886|YNR2_YEAST HYPOTHETICAL 196.1 KD PROTEIN IN RPS3-PSD1 INTERGENIC REGION

gi|2132761|pir||S63127 probable membrane protein YNL172w - yeast (Saccharomyces cerevisiae)

gi|1302144|emb|CAA96060.1| (Z71448) ORF YNL172w [Saccharomyces cerevisiae]

```
Sample ID : Band 1,        HCCA Nan    [Pass:0]
Measured peptides       : 39
Matched peptides        : 26
Min. sequence coverage: 19%
```

ProFound - Search Result Summary

Version 4.10.10
© 1997-2001 ProteoMetrics

Protein Candidates for search 20020305141240-097C-1520190038229 [29120 sequences searched]

Rank	Expectation	Protein Information and Sequence Analyse Tools (T)	%	pI	kDa	®
1	1.5×10^{-16}	gi|6324157|ref|NP_014227.1| (NC_001146) anaphase-promoting complex component; Apc1p [Saccharomyces cerevisiae]	19	5.7	196.13	®
2	0.49	gi|7492952|pir||T40643 probable serine threonine-protein kinase - fission yeast (Schizosaccharomyces pombe)	6	9.5	105.88	®
+3	0.91	gi|18476091|gb|AAK14772.1| (AY026035) septin-1 [Coccidioides immitis]	10	5.1	43.85	®
	-	gi|13398364|gb|AAK21867.1| (AF299320) septin [Aspergillus nidulans]	6	5.0	43.09	®

Figure 1.
Protein identification using peptide mass fingerprinting. The gel plug containing the protein was digested with trypsin, and the peptides were extracted. A sequence coverage of 19% was obtained, allowing the identification of this protein from a database search. Reprinted from [24], with permission.

3.2 Peptide mass fingerprinting

If the sample that was digested was pure, i.e., if it was generated from the digestion of a single protein, the observed result would be a unique signature or fingerprint of that protein. Protein identification of single proteins or simple mixtures can be done by determining the molecular weights (MWs) of the resulting tryptic fragment ions (if trypsin is the enzyme used to digest the protein) and comparing them to a library of such signatures. This approach is called peptide mass fingerprinting [25].

3.3 Sequence-tag approach

More complex mixtures require additional separation steps. In about 2000, a method was developed using two dimensions of mass spectrometry – called MS/MS or tandem mass spectrometry. This is ideal for the analysis of peptides because a peptide cleaves between adjacent amino acids. In this technique (called the sequence-tag approach), a peptide is selected based on its mass in the first stage of mass spectrometry (MS1), fragmented further (in MS2), and then mass analyzed again (in MS3). The amino acid sequence can be manually "read" from the mass differences in the spectrum or the peptide. The peptide molecular weight, and a few fragment ions are often sufficient to identify a protein (**Figure 2**). In the automated version of this approach, the peptide and its parent protein can be identified from a genome-based library of protein amino acid sequences [26–29].

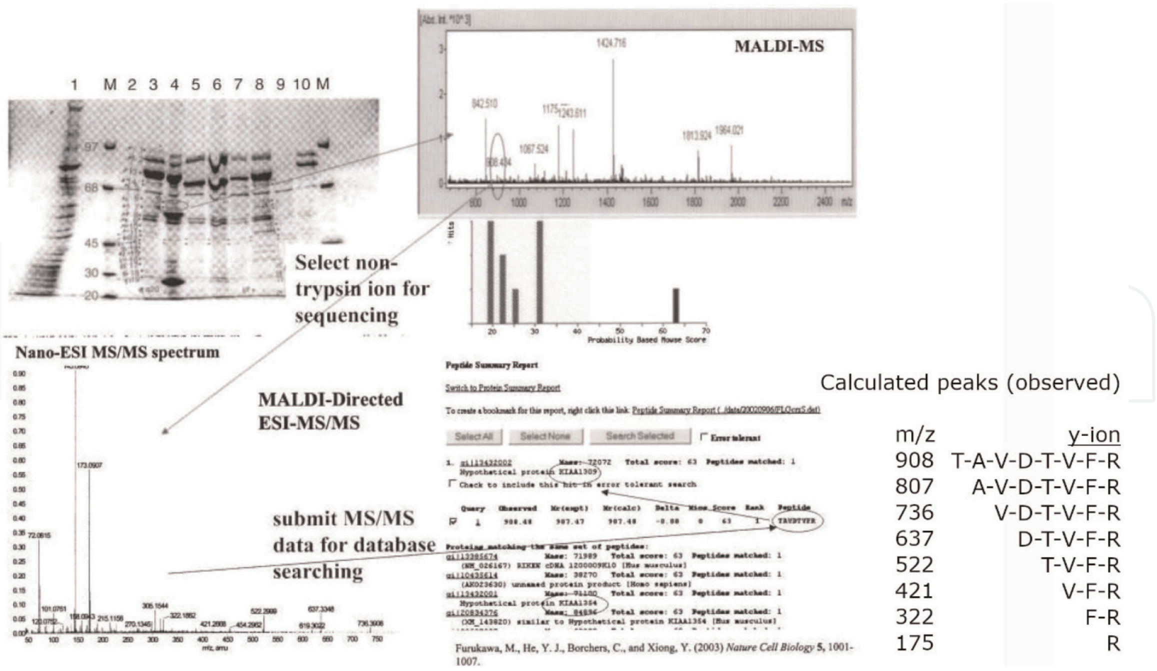


Figure 2.
Protein identification using the sequence-tag approach. Identification of a tryptic digest of a gel-separated protein using the sequence-tag approach. The molecular ion of an unknown peptide (at m/z 908) was selected and fragmented. A nearly complete amino-acid sequence of the peptide T-A-V-D-T-V-F-R was obtained, which led to the identification of the protein KAA1309/KAA1354. Reprinted from [24], with permission.

3.4 Fingerprinting by AI

The application of AI can vastly improve the untangling the complexity of MS-based data analytics. Several database-search software programs are already in widespread use for the interpretation of LC/MS/MS data. These include Mascot [30], Sequest [31], and X!Tandem [32].

Mass spectral peaks can be characterized using their position on the horizontal axis, width, and height. Therefore, a full spectrum $s(x)$ is the summation of several Gaussian distributions, one for each peak,

$$s(x) = \sum_{i=1}^{N} H_i \, exp\left(\frac{-(x-\mu_i)^2}{2\sigma_i{}^2}\right) + \varepsilon(x) \tag{1}$$

where H_i is the peak height, μ_i is the peak position, σ_i is the line width, N is the number of peaks present in the spectrum, and $\varepsilon(x)$ is the noise. With this approximation, we readily obtain N tuples of (H_i, μ_i, σ_i) that characterize the spectrum. If these peaks correspond to peptides, this is the "peptide mass fingerprint" of the protein in the sample. This can now be compared to a library of M reference spectra $\gamma_j(x)$ for a library index j. In fact, we look for the best linear combination of reference spectra that explains the spectrum at hand, which is the optimization problem

$$\min_{\alpha_j}\left(s(x) - \sum_{j=1}^{M} \alpha_j \gamma_j(x)\right) \tag{2}$$

Numerically, we will want to add the additional criterion that this should be a sparse fit, i.e., as many α_j as possible should be equal to zero. As every $\gamma_j(x)$ uniquely corresponds to a specific protein, this results in an identification of the proteins present, with α_j providing the relative abundance of each.

3.5 Temporal evolution of spectra

When a MS analysis is repeated at regular intervals in time leading to a time-series analysis of spectra, i.e., $s(x) \rightarrow s(x, t)$, changes over time can be detected. This is especially interesting as molecular effects in a protein's environment can induce structural changes that imply functional changes in that protein. Mass spectrometry can detect such changes [33]. We have been able to observe changes due to protein degradation as a function of time, and therefore recommend procedures to avoid adverse conditions that could affect clinical results. In **Figure 3**, we display such a time-series which we used to detect protein degradation during processing delays that occur while a plasma sample is waiting to be processed in a clinical setting [34] or during transport at room temperature (**Figure 3**) [35].

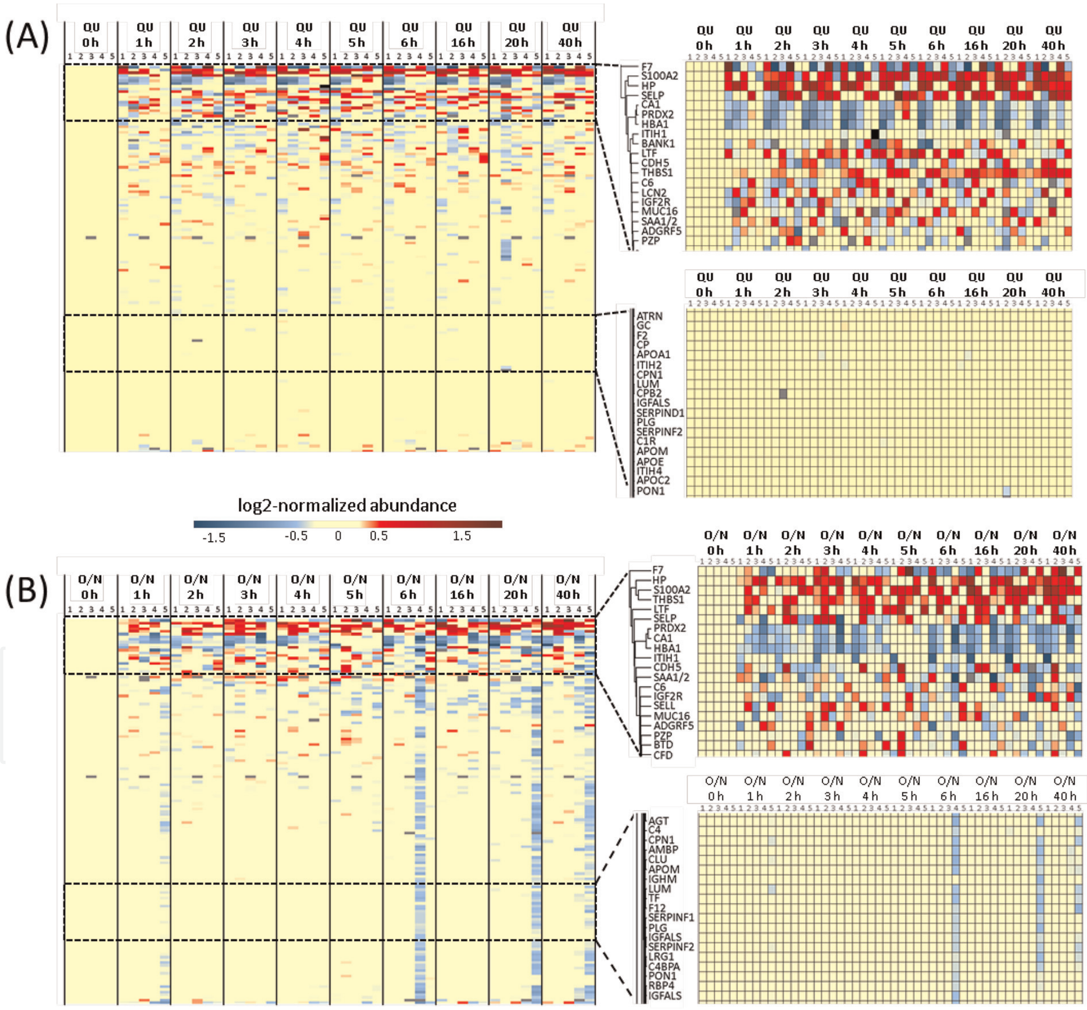


Figure 3.
Heatmap representing 143 proteins quantified in at least 80% of the 50 plasma samples thawed (a) quickly on ice (QU), or (B) overnight at 4°C (O/N). All values were normalized to the respective time point 0 h. missing values are depicted in gray. Yellow areas represent log2 values from −0.25 to 0.25, corresponding to an approximately 1.19-fold difference in protein concentration compared to the respective time point 0 h. blue and red colors indicate concentrations that are >1.2-fold lower/higher than the 0 h time point. Sections of the heatmaps with MRM assays that are unstable or particularly stable are shown, and gene names represented by the surrogate peptides quantified are given. (reprinted from Gaithers, et al. 2022 [34], with permission).

AI has specific impact on MS based proteomics, wherein it facilitates the assignment of fragmentation spectra to actual peptide identifications, and thereby assisting with the quantitation. This data integration has an influence across wide range of proteomics applications including; expression proteomics, networks highlighting complex protein interactions in cellular pathways, impact of post-translational modifications (PTMs), and cross-functional relationships with other omics studies. PTMs of proteins act as signals for localization, degradation, and other functions. Analyzing PTM-bearing peptides is challenging considering the wide spectrum of isoforms, and the complexity of their fragmentation spectra. They might not even follow the tryptic peptide cleavage rules. AI applications can be adopted to solve this challenge and the initial work was highlighted by Wang et al. in 2017 [36], and Yang et al. in 2021 [37]. This can also be applied to Major Histo Compatibility (MHC) Class I and II peptides. The identification rates of these peptides were increased by several fold, as highlighted by Li et al. in 2020 [38].

In addition to direct spectral analysis using DL, the application can also be extended for predicting chromatographic separation retention times. This can be accomplished by direct modeling or through transfer learning [39]. Combining DL for predicting MS/MS spectra with retention times provides more value than experimental library search using Data Independent Acquisition (DIA) [40]. This in turn opens the door for combining Dl with real-time data acquisition. The image captioning feature in DL can be used to sequence peptides from MS/MS without the need for a database (DeepNovo) and has shown a 5–20% improvement over other DeNovo sequencing tools [41]. The DIA-NN (Data Independent Acquisition Neural networks) program is one example where deep neural networks were able to increase the sensitivity of matching DIA fragmentation to *in-silico* derived spectral libraries [42].

There are challenges to MS based analysis of proteins and this pertains to accurately recognizing the protein patterns. MS analysis is done by reviewing smaller parts of amino acid sequences, and then this data is searched against a database, and assigned to proteins. In doing so, some proteins are missed. This challenge can be overcome by training neural networks to recognize protein patterns, with minimal or no errors. Commercial software such as Prosit from the technical University of Munich is available to accomplish this task. AI can also be used to analyze large datasets, and Ml algorithms have been developed by scientists from Novo Nordisk Foundation for Protein Research, and Niels Bohr Institute for recognizing protein patterns in a rapid manner [43].

3.6 Analysis of mass spectra

Much of the data generated by "shotgun" (non-targeted) proteomics is predictable by ML if adequate training data is available. ML methods can analyze the patterns in the data to identify biomarkers, and Dl can be used to extract additional features. Over the past two decades, AI has been increasingly applied to bottom-up MS-based proteomics analyses. In this regard, ML algorithms, have been used for peptide identification, predicting retention times [44–46], as well as for MS/MS spectrum prediction [5, 47, 48]. The earliest ML tool for proteomics was the Percolator algorithm [7], which was able to increase the number of identified peptides by 5–16% [49]. These algorithms has found value in identifying peptides, and have also been used for matching MS1-identified features with value for protein quantitation [50]. It was Zhou, et al. in 2017 [51] who used DL for predicting fragmentation spectra of a peptide from its sequence. DL approaches have gained much value for predicting peptide behavior from sequences, considering the availability of large amount of peptides in databases such as PRIDE [52]. In 2020, Liu, et al. [53] showed that DL can be used to predict full MS/MS spectra, as opposed to just backbone fragment intensities.

In comparing experimental results from one sample to a library of reference data that have themselves been obtained from experiments, we must be aware that measurement errors and noise are ubiquitous. Both sides of the comparison are not just imprecise, but some peptides might be missing in the data – peptides expected by the library might be missing in our data sample, or peptides present in our data sample might be missing in the library [54]. In practice, the matching is often done via a similarity score [21]. Despite this development, the key challenge for mass spectrometry in proteomics is to develop data analysis tools [8].

In some of the established analysis tool, outliers are removed by performing the Grubbs test [55], but the foundational assumption of this test is that the data is normally distributed. In the case of a normal spectrum, this is unfortunately not the case, as most data points are at the low end of the relative abundance scale. We must find other ways of cleaning up a spectrum and analyzing it.

Depending on the technique, the peaks in the mass spectrum will either be individual peptides (in peptide mass fingerprinting), or fragment ions (in a tandem MS (i.e., an MS/MS) experiment). The peak heights or intensities depend on the relative sensitivities of the peptide or fragment ion and the abundances of that component in the mixture under investigation. As with all experimental methods, the presence of noise makes analysis difficult and can hide small peaks in the spectrum; the well-known signal-to-noise (S/N) ratio challenge. Similarly, this analysis depends on the availability of a library of known spectra for proteins or peptides – standards against which the experimental spectrum must be compared. The scientific analysis of the data must begin with an experimental spectrum by removing as much noise from it as possible. There are two common ways of doing this: filtering and subtraction.

One method of filtering is to transform the spectrum into its companion domain by using the Fourier Transform. We may then delete the low frequency (high-band-pass filter) or high frequency (low-band-pass filter) features. We can also multiply the frequency features by a factor (Wiener filter [56]) in order to suppress the noisy elements not caught by band passing. If we have a more sophisticated idea of the mechanism producing the noise, we could use a Kalman filter [57] as well, which is a statistical estimation of the joint probability distribution of all independent variables in an effort to extract a linear quadratic probabilistic model of the error. After this, the spectrum is transformed back into its original form, again using a Fourier Transform. It is now de-noised. These methods are very computationally efficient and have been incorporated into in many different software tools. If the noise source is deemed to be white noise (i.e., structure free), then these methods are probably the best one can do.

Subtraction is a method by which the noise signal is simply subtracted from the real signal. For this, we must either measure the noise signal or model it in some computational way. The way to experimentally measure the noise is to pass a well-known pure protein source through the identical experimental setup and measure its signal. This would best be done with several different pure proteins so that the signals could be compared, and a holistic noise spectrum built up. Modeling the noise source amounts to constructing a first-principles computer simulation of the process, which may not be commercially feasible.

Unfortunately, real-world noise is always structured and never completely white. Thus, removing structured noise consumes significantly more effort. This is the reason that so much developmental work in mass spectrometry instrumentation has been focused on the physical apparatus in trying to get ever cleaner signals [58]. Nonetheless, these can always be cleaned further by data analytics.

4. Applications

In living systems, proteins can exert their functional influence either individually and/or by interacting with other molecules. The latter can be studied in-depth by using interactomics to pull down the complexes, and then analyzing the network of molecules involved in the interactions. The data derived from this analysis needs to be integrated and studied along with data from genomics, epigenomics, transcriptomics, metabolomics, and lipidomics. AI is gaining increased attention both for biomarker discovery, and validation, as well as for integrating proteomics data with other omics data along with clinical data to derive a more complete picture of the patient's health status. Specifically the data can be used in predicting 3D structure of a target protein, drug-protein interactions, and *de novo* drug design.

4.1 Biomarker discovery and validation

MS based proteomics can be leveraged to identify the "needle in the haystack" along with quantitative data, with one or more unique proteins turning out to be the biomarkers of choice for clinical diagnostics, and/or drug discovery – potentially leading to development of therapeutics. Commencing the analysis with a large sample set from multiple cohorts, proteomic datasets that can be assembled into correlation maps [59]. This will allow us to identify and understand proteins that are co-regulated and those that are connected in their functional pathways. ML can aid in uncovering signals and predictors, leading to the creation of decision trees, and grouping classes based on specific features. Methods are also available to differentiate between classes based on unique features, and it assists in interpretability, and accurate classification [60]. One pitfall to watch out for is the over-fitting arising from the training dataset. Identification of biomarkers by ML needs rigorous oversight on classifying the positives, and negatives diagonally, with the incorrect ones being off-diagonal, and reflected in the ROC curves.

Practical outcomes for applying ML techniques to proteomic data sets for circulating biomarkers in alcoholic liver disease was shown by Matthias Mann's group [61]. Additionally, they also showed the classification of biomarkers for Alzheimer's disease from CSF in a three cohort study [62]. In their observations, they felt that the ML methods performed well, with the only consideration being the data points used to train the model. The next step would be to translate these potential biomarkers into a clinical test, which will undergo a rigorous scrutiny by the regulatory agencies. If this involves a multi-analyte panel, then the entire pipeline, including the ML method needs to be simple, robust, explainable, and reproducible in order to obtain the regulatory approval. In this regard the ML acts as the final step in filtering the complex proteomics dataset into selectable biomarker candidates that can pass the acid test of the regulatory bodies. One of the key considerations for the outcomes from the AI based analysis is the chance of bias when the datasets are derived from a non-diverse source [63]. Overcoming this bias will require the analysis of samples from diverse populations.

Although a few proteins in a cell are present in high abundance, it is the remaining tens of thousands of proteins that are potentially interesting as biomarkers [17–20]. To determine a biomarker of a disease, one has to be able to quantitatively compare samples from patients with different disease states. This poses additional challenges, the first is detection sensitivity, the second (if the method is to have clinical value), it has to be reproducible in both intra- and inter laboratory studies [64, 65].

AI has been successfully applied to biomarker identification based on MS data [10]. In a recent paper [66], machine learning was used to accurately predict the survival of COVID patients on the day they were admitted to the hospital ICU, based on the plasma concentrations of 10 proteins and 5 metabolites, as determined by MS. While humans can visually detect differences in the concentration of a single protein biomarker, but for sorting out the data from thousands of proteins, computerized methods are needed. In effect, one is trying to detect a decision surface in n-dimensional space. This is typically done by Principal Component Analysis (PCA). Both the proteomics and metabolomics markers allowed the prediction of survival separately, with accuracies of 83% (AUC 0.90) and 84% (AUC 0.93), respectively. When combined, however, the concentration measurements of all markers yielded an accuracy of 92% (AUC 0.97), see **Figure 4**.

When combined with real-world evidence (RWE) about diseases [67, 68] from published literature, databases, images, patient health records, and clinical data [69, 70] the overall insight lends itself well for the current trends in personalized/precision medicine. From individual to cohorts, this analysis can be extended to population studies, where ML, DL, and text mining lend themselves to the extraction of relationships between all interacting partners [71, 72]. Knowledge graphs [73] highlight the complex interactions in a biological network, and the insight derived from analyzing these knowledge graphs are more insightful. Clinical Knowledge Graphs (CKG) which contain millions of protein nodes, and greater than 100 million connections were developed by Mathias Mann's group [74]. This in turn has been used to integrate knowledge from the proteome project [75], and ML and DL algorithms can be leveraged to study these knowledge graphs.

4.2 Multiomics and proteomic data integration

Deriving a meaningful outcome that is representative of the final phenotype also needs the interaction of other omics data along with that of the proteomics data. Molecular interactions in living systems are complex, and proteins tend to interact with nucleic acids, other proteins, lipids, and small molecules in exerting the cellular function. Proteins bind to DNA or RNA which, in turn, has an influence on replication, transcription, DNA repair, and transport, translation, splicing, and silencing in the case of RNA [76]. Protein-protein interactions are critical in many cellular processes; for example in the use of therapeutic monoclonal antibodies [77], peptide drugs [78], and chimeric antigen receptor T cells [79]. Proteins interact with lipids in multiple ways and this has functional significance such as selective transport, and cell-cell communication [80]. Finally, protein interactions with small molecules especially drugs can be analyzed by leveraging mass spectrometry [81].

Elucidating these complex interactions of proteins with other molecules can be accomplished by using baits in a pull-down assay [82]. These interactome analyses generates complex data, and the application of ML, and DL technologies is still at the early stages with most of it being used for data pre-processing [83]. As of today, the lack of detailed analyses using AI techniques probably arises from the lack of adequate

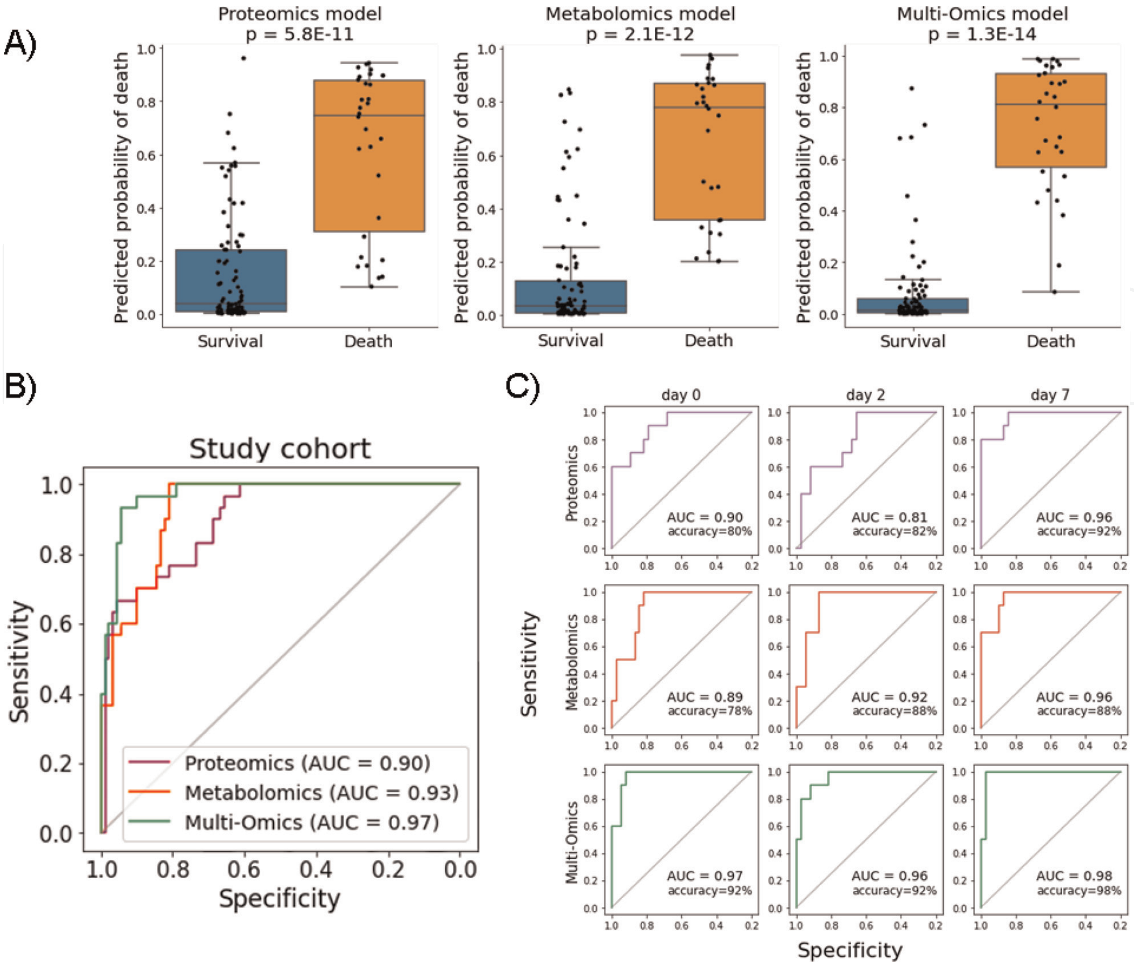


Figure 4.
Reliable and accurate prediction of survival upon hospitalization. (A) The performance of the support vector machine classifier was used to predict COVID-19 patient survival based on proteomics (10 proteins), metabolomics (5 metabolites), and combined multi-omics models (10 proteins +5 metabolites) and using all data points (days 0, 2, 7 after admission). (B) Receiver-operating characteristic (ROC) curves show that the best performance was obtained with the multi-omics model (10 proteins +5 metabolites). (C) ROC curve analysis for proteomics-only, metabolomics-only, and multi-omics models at different time points after admission (days 0, 2, or 7). Upper row – Proteomics model based on 10 proteins, middle row – Metabolomics model based on 5 metabolites, bottom row – Combined multi-omics model based on 10 proteins and 5 metabolites. (figure and figure legend from [66], with permission).

sample sets of data, validation data, and the ability to interpret the data [63]. Ribosomal Binding Protein-RNA (RBP-RNA) interactions are of importance to drug discovery and to other biological functions, because of their critical role in gene expression [84]. The paradigms for studying these interactions include; binding site prediction, binding preference prediction, and a merged approach that predicts both the binding sites, and binding preference [85]. DL methodologies add value to the study of RBP-RNA interactions by operating directly on the structure data [86, 87].

Proteins tend to operate through complexes to maintain the functional aspects of a living organism. Understanding the interactome, will intern shed light on the complex aspects of biology in these organisms, and also facilitate drug development. Lack of structural knowledge of many proteins prevented the construction of an interactome. This changed in 2020 when DeepMind released AI technologies termed "AlphaFold", and "RoseTTAFold", that enabled the prediction of protein structures from the gene sequences. Researchers from UT Southwestern and University of Washington, used AI and evolutionary analysis to produce 3D models of eukaryotic protein interactions.

Their study, published in Science ([88], identified 100 protein complexes, and provided structural models for 700 previously uncharacterized ones [89].

It has been estimated that 25% of cellular proteins are membrane proteins [90] that can be either transmembrane or positioned on the periphery [91]. Cellular membranes are one of the fundamental protective elements of the cell wherein the proteins are embedded within the lipid bilayer. Besides offering structural integrity, these proteins also have a functional role ranging from molecular transport, signal transduction, cell-cell recognition, attachment, and enzymatic activity. Molecular dynamics (MD) simulations have been used to understand the lipid-protein interactions, as well as the structure and dynamics as influenced by the proteins in the membranes [80]. Ion channels are an example of protein-lipid structures formed in the membranes, and they are directly or indirectly associated with cellular disorders indicative of certain diseases. Hence these ion channels are used as targets for drug discovery and therapy. AI techniques have proven to be of value in predicting the related genes, mutations, and the relationship to certain diseases. Taju et al. has used DL methods for the classification of ion transporters, and ion channels from membrane proteins, by training the deep neural networks using the position-specific scoring matrix profile as the input [92]. ML has been used to derive the feature vectors of ion channels including SVMProt, and k-skip-n-gram methods, 188-, and 400 dimensional features, respectively [93].

4.3 Structural proteomics

In addition to its value for biomarker discovery, and validation (Section 4.1), and proteomic data integration from multiomics analysis (Section 4.2), AI also offers immense value in structural proteomics. This arises from the fact that proteins have 4 levels of structure, namely primary, secondary, tertiary, and quaternary. The complexity of the structure increases as the proteins assumes the tertiary, and quaternary structure, and associates with other molecules. A typical protein contains 200–400 amino acids (**Figure 5**) and typically folds into a three-dimensional shape. It has been estimated that the number of possible protein folds found in nature is anywhere from a few hundred [97] to approximately 2200 [98] different folds. Protein misfolding leading to aggregation of the prion protein is the cause of Creutzfeldt-Jakob disease (i.e., “mad cow disease” in animals) and of cystic fibrosis, most often caused by one missing amino acid (F508Δ) out of the ∼1480 amino acids in the normal CFTR protein. Protein misfolding may also be the involved in diseases such as Alzheimer’s or Huntington’s disease. On the other hand, proteins such as antibodies can also be the cure for viral and bacterial infections.

Molecular modeling and molecular dynamics simulations have already played a large part in protein structure determination [99–103]. The breakthrough Alphafold [104, 105] program will undoubtedly also play a large role in protein structure determination, although there are limitations on its applicability, particularly for dynamic systems and stems where there can be multiple states of a protein [106, 107].

Crosslinking has played a key role in protein structure determination, both for individual proteins and protein complexes. Crosslinks of different lengths, connecting specific amino acid residues, and can be used as a molecular ruler [108] to determine the distances between these residues in a folded protein. Crosslinks connecting residues that are too far apart in a structural model crosslinker can be used to “rule out” certain structures (pun intended). Because a mass spectrum is essentially a plot of intensity vs. mass, crosslinked peptides can be detected because they have higher

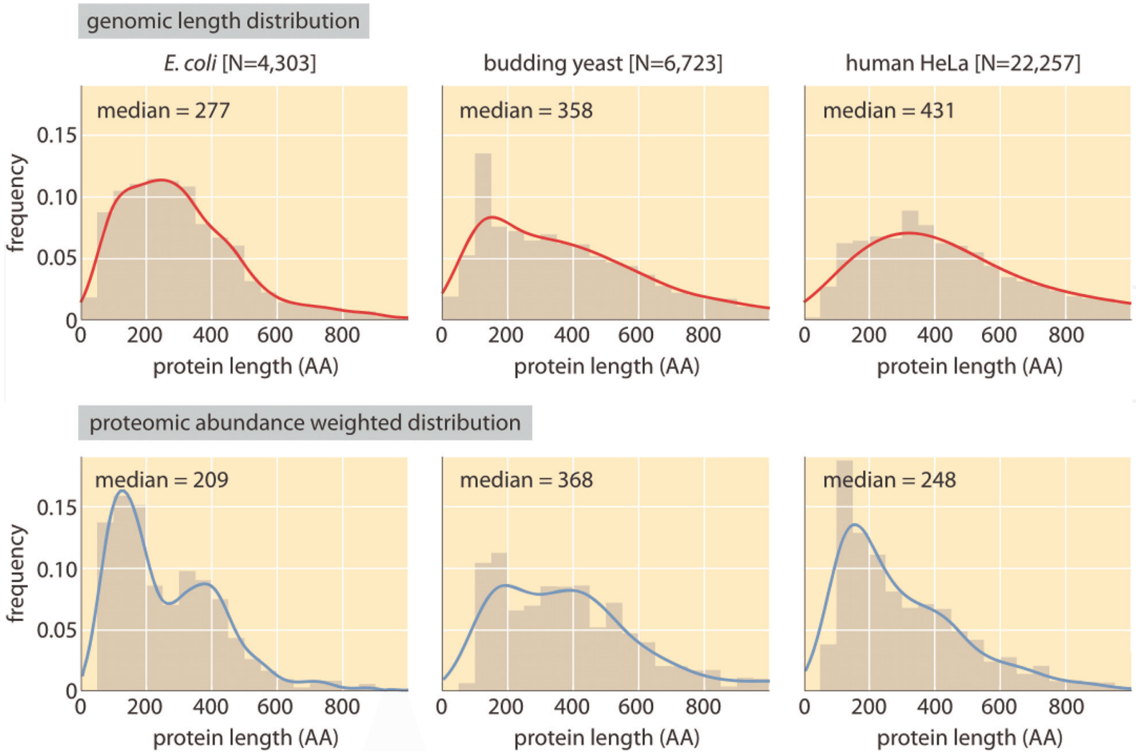


Figure 5.
Distribution of protein lengths in Escherichia coli, *budding yeast and human HeLa cells. (TOP) protein length is calculated in amino acids (AA), based on the coding sequences in the genome. (BOTTOM) distributions are drawn after weighting each gene with the protein copy number inferred from mass spectrometry proteomic studies [94], M9 + glucose [95, 96]. Continuous lines are Gaussian kernel-density estimates for the distributions serving as a guide to the eye. (figure and figure legend from [11], http://book.bionumbers.org/how-big-is-the-average-protein/#:~:text=A%20simple%20rule%20of%20thumb,proteins%20that%20make%20cells%20work.*

molecular weights than unlinked residues. Crosslinkers have been specifically designed for use with MS-based research [109–111]. Features included in these crosslinkers include the use of stable isotopes to create unique and detectable signatures (patterns) in the mass spectrum that can be detected by software programs [110].

In addition to providing confirmation of MD results, MS-derived crosslinking data has been used as constraints in molecular dynamics simulations [112], with the result of reducing the conformational space and achieving the correct protein folding on practical time scales [113]. We have applied this approach to structural studies of prion monomers and β-oligomers [114], as well as native α-synuclein [115] and the "unstructured" Tau protein [116].

5. Challenges of artificial intelligence

Applying AI to proteomics is challenging for some reasons that are unique to proteomics [117] or medicine [118], and also for some reasons that are general [119]. The most important of these challenges are as follows. First, expertise in both AI and proteomics is required and this is rarely available in the same individual and thus requires interdisciplinary teamwork [120]. Second, data must generally be labeled or annotated by human experts to prepare the data for AI model training, which is a time-consuming and expensive activity that dominates the total cost of the project

[121]. Third, models trained on readily available, small, or restricted datasets often fail to generalize to real-life datasets and thus do not deliver value [122]. Fourth, AI cannot be a "black box" and so either the model must be interpretable [123], or the model's output explainable [124] to provide actionable insights to a human actor. Fifth, it is a misconception that AI is rapid and inexpensive because it is automated – in fact, doing responsible and sustainable AI requires a long-term investment in people and infrastructure.

Another challenge specific to proteomics, and biomarker discovery is the overfitting of the training data for ML and DL data [63], where the model is biased towards the information in the training data, leaving out the information in new and unseen data. This could lead to differences in biomarkers identified in different patient cohorts. This challenge could be overcome by leveraging the largest training and testing data, followed by cross-validation on both the training, and unseen data.

AI is not a specific technology but rather an umbrella term for many different methods. The overriding idea is the creation of a mathematical summary of empirical data that can be used to interpolate and extrapolate beyond the empirical data to make predictions about new situations. The dividing line between AI and related disciplines, such as statistics, can get blurred.

6. Conclusion and future outlook

AI represents a powerful toolbox that represents substantial value for proteomics and life-sciences in general. There are many applications already, as we have seen, and many more will arise in the near future. These are necessarily inter-disciplinary efforts that require experts from both AI and the domain. Notwithstanding the above mentioned challenges, the work for AI is dominated by preparing a suitable, large, clean, representative, and significant dataset. This will make AI's promise more accessible to those who already have large amounts of data and the experts to prepare such data for analysis.

The value of AI lies in automation and in making processes accessible in the first place through its speed and automation. The most promising such direction is personalized medicine. If a single individual's proteome could be inferred from a blood sample in an instant, the troublesome proteins or the absent proteins could be identified, and suitable strategies developed for producing counter-agents or the missing links, this could produce a treatment specific to that person as opposed to giving everyone with some condition the same mass-produced treatment and hoping for the best.

Author details

Patrick Bangert[1*], Nandha Kumar Balasubramaniam[2*], Carol E. Parker[3]
and Christoph H. Borchers[3,4,5]

1 Samsung SDS, San Jose, California, USA

2 Oracle, Redwood City, CA, USA

3 Segal Cancer Proteomics Centre, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, Quebec, Canada

4 Gerald Bronfman Department of Oncology, Division of Experimental Medicine, Lady Davis Institute for Medical Research, McGill University Faculty, Montreal, Quebec, Canada

5 Department of Pathology, Jewish General Hospital, McGill University Faculty of Medicine and Health Sciences, Montreal, Quebec, Canada

*Address all correspondence to: p.bangert@samsung.com and ceekpeace@gmail.com

References

[1] Hassabis D. Alpha Fold Reveals the Structure of the Protein Universe. Deepmind. 2022 https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe.

[2] HUPO. A gene-centric human proteome project: HUPO–the human proteome organization. Molecular & Cellular Proteomics. 2010; **9**:427-429

[3] Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2007;**104**: 19428-19433

[4] Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, et al. The human proteome project: Current state and future direction. Molecular & Cellular Proteomics. 2011;**10**: M111.00993

[5] Arnold RJ, Jayasankar N, Aggarwal D, Tang H, Radivojac P. A machine learning approach to predicting peptide fragmentation spectra. Pacific Symposium on Biocomputing. 2006;**11**: 219-230

[6] Gabriels R, Martens L, Degroeve S. Updated MS^{2}PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. Nucleic Acids Research. 2019;**47**:W295-W299

[7] Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods. 2007;**4**: 923-925

[8] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;**422**:198-207

[9] McLafferty FW, Breuker K, Jin M, Han X, Infusini G, Jiang H, et al. Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. The FEBS Journal. 2007;**274**:6256-6268

[10] Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: Classification and biomarker identification in postgenomics biology. Omics: a Journal of Integrative Biology. 2013;**17**:595-610

[11] Milo R, and Phillips R. Cell Biology by the Numbers. How big is the average protein? Taylor and Francis. 2015 http://book.bionumbers.org/how-big-is-the-average-protein/#:~:text=A%20simple%20rule%20of%20thumb,proteins%20that%20make%20cells%20work

[12] Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, et al. Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. Biotechnology & Genetic Engineering Reviews. 1996;**13**:19-50

[13] Schiess R, Wollscheid B, Aebersold R. Targeted proteomic strategy for clinical biomarker discovery. Molecular Oncology. 2009;**3**:33-44

[14] Mayeux R. Biomarkers: Potential uses and limitations. NeuroRx: the journal of the American Society for Experimental NeuroTherapeutics. 2004; **1**:182-188

[15] Kohler S. Precision medicine – Moving away from one-size-fits-all.

Quest -Science for South Africa. 2018;**14**: 12-15

[16] Anderson NL, Anderson NG. The human plasma proteome: History, character, and diagnostic prospects. Molecular & Cellular Proteomics. 2002; **1**:845-867

[17] Baker ES, Liu T, Petyuk VA, Burnum-Johnson KE, Ibrahim YM, Anderson GA, et al. Mass spectrometry for translational proteomics: Progress and clinical implications. Genome Medicine. 2012;**4**:63-73

[18] Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. Nature Biotechnology. 2006;**24**:971-983

[19] Anderson NL. The roles of multiple proteomic platforms in a pipeline for new diagnostics. Molecular & Cellular Proteomics. 2005;**4**:1441-1444

[20] Jacobs JM, Adkins JN, Qian WJ, Liu T, Shen Y, Camp, D. G. n., and Smith, R. D. Utilizing human blood plasma for proteomic biomarker discovery. Journal of Proteome Research. 2005;**4**:1073-1085

[21] Carr S. Fundamentals of Biological Mass Spectrometry and Proteomics. Cambridge, MA: Broad Institute

[22] Griffin TJ, Aebersold R. Advances in proteome analysis by mass spectrometry. The Journal of Biological Chemistry. 2001;**276**:45497-45500

[23] Cañas B, López-Ferrer D, Ramos-Fernández A, Camafeita E, Calvo E. Mass spectrometry technologies for proteomics. Briefings in Functional Genomics & Proteomics. 2006;**4**:295-320

[24] Parker CE, Warren MR, Loiselle DR, Dicheva NN, Scarlett CO, Borchers CH. Identification of components of protein complexes. Methods in Molecular Biology. 2005;**301**:117-151

[25] Cottrell JS. Protein identification by peptide mass fingerprinting. Peptide Research. 1994;**7**:115-124

[26] Eng JK, McCormack AL, Yates JR. 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. 1994;**5**:976-989

[27] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Analytical Chemistry. 1994;**66**: 4390-4399

[28] Mortz E, Vorm O, Mann M, Roepstorff P. Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. Biological Mass Spectrometry. 1994;**23**: 249-261

[29] Mortz E, O'Connor PB, Roepstorff P, Kelleher NL, Wood TD, McLafferty FW, et al. Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. Proceedings of the National Academy of Sciences of the United States of America. 1996;**93**:8264-8267

[30] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;**20**:3551-3567

[31] Yates JR, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Analytical Chemistry. 1995;**67**: 1426-1436

[32] Fenyo D, Eriksson J, Beavis R. Mass spectrometric protein identification using the global proteome machine. Methods in Molecular Biology. 2010;**673**: 189-202. DOI: 10.1007/978-1-60761-842-3_11

[33] Cappelletti V, Hauser T, Piazza I, Pepelnjak M, Malinovska L, Fuhrer T, et al. Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. Cell. 2021;**184**:545-559. e522

[34] Gaither C, Popp R, Zahedi RP, Borchers CH. Multiple reaction monitoring-mass spectrometry enables robust quantitation of plasma proteins regardless of whole blood processing delays that may occur in the clinic. Molecular & Cellular Proteomics. 2022; **21**:100212

[35] Gaither C, Popp R, Borchers SP, Skarphedinsson K, Eiriksson FF, Thorsteinsdóttir M, et al. Performance assessment of a 125 human plasma peptide mixture stored at room temperature for multiple reaction monitoring-mass spectrometry. Journal of Proteome Research. 2021;**20**: 4292-4302

[36] Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, et al. MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics. 2017;**33**: 3909-3916

[37] Yang Y, Horvatovich P, Qiao L. Fragment mass Spectrum prediction facilitates site localization of phosphorylation. Journal of Proteome Research. 2021;**20**:634-644

[38] Li K, Jain A, Malovannaya A, Wen B, Zhang B. DeepRescore: Leveraging deep learning to improve peptide identification in Immunopeptidomics. Proteomics. 2020;**20**:e1900334

[39] Ma C, Ren Y, Yang J, Ren Z, Yang H, Liu S. Improved peptide retention time prediction in liquid chromatography through deep learning. Analytical Chemistry. 2018;**90**:10881-10888

[40] Yang Y, Liu X, Shen C, Lin Y, Yang P, Qiao L. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nature Communications. 2020;**11**:146

[41] Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. Proceedings of the National Academy of Sciences of the United States of America. 2017;**114**: 8247-8252

[42] Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. Nature Methods. 2020;**17**:41-44

[43] Goled S. How Artificial Intelligence Is Reviving Proteomics. 2022 https://analyticsindiamag.com/how-artificial-intelligence-is-reviving-proteomics/#:~:text=Recent%20Developments%20in%20Use%20Of%20AI%2FML%20In%20Proteomics&text=It%20analyses%20smaller%20parts%20consisting,and%20assigned%20to%20specific%20proteins

[44] Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O. Statistical learning of peptide retention behavior in chromatographic separations: A new kernel-based approach for computational proteomics. BMC Bioinformatics. 2007;**8**:468

[45] Moruz L, Käll L. Peptide retention time prediction. Mass Spectrometry Reviews. 2017;**36**:615-623

[46] Moruz L, Tomazela D, Käll L. Training, selection, and robust calibration of retention time models for targeted proteomics. Journal of Proteome Research. 2010;**9**: 5209-5216

[47] Degroeve S, Martens L. MS2PIP: A tool for MS/MS peak intensity prediction. Bioinformatics. 2013;**29**: 3199-3203

[48] Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nature Biotechnology. 2004;**22**: 214-219

[49] Granholm V, Kim S, Navarro JC, Sjölund E, Smith RD, Käll L. Fast and accurate database searches with MS-GF+percolator. Journal of Proteome Research. 2014;**13**:890-897

[50] The M, Käll L. Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. Nature Communications. 2020;**11**:3234

[51] Zhou XX, Zeng WF, Chi H, Luo C, Liu C, Zhan J, et al. pDeep: Predicting MS/MS spectra of peptides with deep learning. Analytical Chemistry. 2017;**89**: 12690-12697

[52] Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Research. 2019;**47**: D442-D450

[53] Liu K, Li S, Wang L, Ye Y, Tang H. Full-Spectrum prediction of peptides tandem mass spectra using deep neural network. Analytical Chemistry. 2020;**92**: 4275-4283

[54] Aebersold R, Goodlett DR. Mass spectrometry in proteomics. Chemical Reviews. 2001;**101**:269-295

[55] Park SK, Venable JD, Xu T, Yates JRI. A quantitative analysis software tool for mass spectrometry–based proteomics. Nature Methods. 2008;**5**:319-322

[56] Rabiner LR, Gold B. Theory and Application of Digital Signal Processing. Hoboken, New Jersey: Prentice Hall; 1975

[57] Kalman AH. Fundamentals of Adaptive Filtering. Hoboken, NJ: John Wiley & Sons, Inc.; 2003

[58] Timp W, Timp G. Beyond mass spectrometry, the next step in proteomics. Science Advances. 2020;**6**: eaax8978

[59] Wewer Albrechtsen NJ, Geyer PE, Doll S, Treit PV, Bojsen-Møller KN, Martinussen C, et al. Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after roux-En-Y gastric bypass surgery. Cell Systems. 2018;**7**:601-612.e613

[60] Geyer PE, Voytik E, Treit PV, Doll S, Kleinhempel A, Niu L, et al. Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies. EMBO Molecular Medicine. 2019;**11**:e10427

[61] Niu L, Thiele M, Geyer PE, Rasmussen DN, Webel HE, Santos A, et al. A paired liver biopsy and plasma proteomics study reveals circulating biomarkers for alcohol-related liver disease. bioRxiv. 2020:2020

[62] Bader JM, Geyer PE, Müller JB, Strauss MT, Koch M, Leypoldt F, et al. Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's

disease. Molecular Systems Biology. 2020;**16**:e9356

[63] Mann M, Kumar C, Zeng WF, Strauss MT. Artificial intelligence for proteomics and biomarker discovery. Cell Systems. 2021;**12**:759-770

[64] Paulovich AG, Whiteaker JR, Hoofnagle AN, Wang P. The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. Proteomics. Clinical Applications. 2008; **2**:1386-1402

[65] Bala K. Healing the Achilles heel of proteomics. Genetic Engineering and Biotechnology News. 1 Feb 2010;**30**(3). Available from: https://www.genengnews.com/magazine/127/healing-the-achilles-heel-of-proteomics

[66] Richard VR, Gaither C, Popp R, Chaplygina D, Brzhozovskiy A, Kononikhin A, et al. Early prediction of COVID-19 patient survival by targeted plasma multi-omics and machine learning. Molecular & Cellular Proteomics. Oct 2022;**21**(10):100277. DOI: 10.1016/j.mcpro.2022.100277. Epub 3 Aug 2022

[67] Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer. 2020;**1**:800-810

[68] Rawshani A, Eliasson B, Rawshani A, Henninger J, Mardinoglu A, Carlsson Å, et al. Adipose tissue morphology, imaging and metabolomics predicting cardiometabolic risk and family history of type 2 diabetes in non-obese men. Scientific Reports. 2020;**10**:9973

[69] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nature Reviews. Genetics. 2012;**13**:395-405

[70] Landi I, Glicksberg BS, Lee HC, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ Digital Medicine. 2020;**3**:96

[71] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: From information retrieval to biological discovery. Nature Reviews. Genetics. 2006;**7**:119-129

[72] Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: Enabling integrative biology. Nature Reviews. Genetics. 2012;**13**:829-839

[73] Callahan TJ, Tripodi IJ, Pielke-Lombardo H, Hunter LE. Knowledge-based biomedical data science. Annual Review of Biomedical Data Science. 2020;**3**:23-41

[74] Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nature Biotechnology. 2022;**40**:692-702

[75] Müller JB, Geyer PE, Colaço AR, Treit PV, Strauss MT, Oroshi M, et al. The proteome landscape of the kingdoms of life. Nature. 2020;**582**:592-596

[76] Cozzolino F, Iacobucci I, Monaco V, Monti M. Protein-DNA/RNA interactions: An overview of investigation methods in the -omics era. Journal of Proteome Research. 2021;**20**: 3018-3030

[77] Weiner LM, Surana R, Wang S. Monoclonal antibodies: Versatile platforms for cancer immunotherapy.

Nature Reviews. Immunology. 2010;**10**: 317-327

[78] Walensky LD, Bird GH. Hydrocarbon-stapled peptides: Principles, practice, and progress. Journal of Medicinal Chemistry. 2014;**57**: 6275-6288

[79] Brentjens RJ, Davila ML, Riviere I, Park J, Wang X, Cowell LG, et al. CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. Science Translational Medicine. 2013; **5**(177):177ra38. DOI: 10.1126/scitranslmed.3005930

[80] Muller MP, Jiang T, Sun C, Lihan M, Pant S, Mahinthichaichan P, et al. Characterization of lipid-protein interactions and lipid-mediated modulation of membrane protein function through molecular simulation. Chemical Reviews. 2019;**119**:6086-6161

[81] Bennett JL, Nguyen G, Donald WA. Protein-small molecule interactions in native mass spectrometry. Chemical Reviews. 2022;**122**:7327-7385

[82] Bludau I, Aebersold R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. Nature Reviews. Molecular Cell Biology. 2020;**21**:327-340

[83] Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S. Deep metabolome: Applications of deep learning in metabolomics. Computational and Structural Biotechnology Journal. 2020; **18**:2818-2825

[84] Dai H, Umarov R, Kuwahara H, Li Y, Song L, Gao X. Sequence2Vec: A novel embedding approach for modeling transcription factor binding affinity landscape. Bioinformatics (Oxford, England). 2017;**33**:3575-3583

[85] Wei J, Chen S, Zong L, Gao X, Li Y. Protein-RNA interaction prediction with deep learning: Structure matters. Briefings in Bioinformatics. 2022;**23**:bbab540

[86] Lam JH, Li Y, Zhu L, Umarov R, Jiang H, Héliou A, et al. A deep learning framework to predict binding preference of rna constituents on protein surface. Nature Communications. 2019;**10**:4941

[87] Li H, Tian S, Li Y, Fang Q, Tan R, Pan Y, et al. Modern deep learning in bioinformatics. Journal of Molecular Cell Biology. 2020;**12**:823-827

[88] Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J. Computed structures of core eukaryotic protein complexes. Science. 2021;**374**:eabm4805

[89] UT_Southwestern_Medical_Center. Artificial intelligence successfully predicts protein interactions. Science Daily. 2021 https://www.sciencedaily.com/releases/2021/11/211116175100.htm#:~:text=Summary%3A,than%20700%20previously%20uncharacterized%20ones.

[90] Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Science. 1998;**7**:1029-1038

[91] von Heijne G, Manoil C. Membrane proteins: From sequence to structure. Protein Engineering. 1990;**4**:109-112

[92] Taju SW, Ou YY. DeepIon: Deep learning approach for classifying ion transporters and ion channels from membrane proteins. Journal of Computational Chemistry. 2019;**40**: 1521-1529

[93] Ashrafuzzaman M. Artificial intelligence, machine learning and deep learning in Ion Channel bioinformatics. Membranes. 2021;**11**:672

[94] Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent *Escherichia coli* proteome. Nature Biotechnology. 2016;**34**:104-110

[95] de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature. 2008;**455**:1251-1254

[96] Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Molecular & Cellular Proteomics. 2012;**11**: M111.014050

[97] Schaeffer RD, Daggett V. Protein folds and protein folding. Protein Engineering, Design & Selection : PEDS. 2011;**24**:11-19

[98] Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins. 1999; **35**:408-414

[99] Petrotchenko EV, Borchers CH. Protein chemistry combined with mass spectrometry for protein structure determination. Chemical Reviews. 2022; **122**:7488-7499

[100] Medina-Franco JL, Méndez-Lucio O, Martinez-Mayorga K. The interplay between molecular modeling and chemoinformatics to characterize protein-ligand and protein-protein interactions landscapes for drug discovery. Advances in Protein Chemistry and Structural Biology. 2014; **96**:1-37

[101] Roel-Touris J, Jiménez-García B, Bonvin A. Integrative modeling of membrane-associated protein assemblies. Nature Communications. 2020;**11**:6210

[102] Soni N, Madhusudhan MS. Computational modeling of protein assemblies. Current Opinion in Structural Biology. 2017;**44**:179-189

[103] Geromichalos GD. Importance of molecular computer modeling in anticancer drug development. Journal of B.U.ON. : official journal of the Balkan Union of Oncology. 2007;**12**:S101-S118

[104] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; **596**:583-589

[105] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Applying and improving AlphaFold at CASP14. Proteins. 2021;**89**:1711-1721

[106] Ruff KM, Pappu RV. AlphaFold and implications for intrinsically disordered proteins. Journal of Molecular Biology. 2021;**433**:167208

[107] Perrakis A, Sixma TK. AI revolutions in biology: The joys and perils of AlphaFold. EMBO Reports. 2021;**22**:e54046

[108] Green NS, Reisler E, Houk KN. Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers. Protein Science. 2001;**10**:1293-1304

[109] Brodie NI, Makepeace KA, Petrotchenko EV, Borchers C. Isotopically-coded short-range hetero-bifunctional photo-reactive crosslinkers for studying protein structure. Journal of Proteomics. 2015;**118**:12-20

[110] Petrotchenko EV, Serpa JJ, Borchers CH. An isotopically-coded CID-cleavable biotinylated crosslinker for structural proteomics. Molecular & Cellular Proteomics. 2011;**10**: M110.001420

[111] Petrotchenko EV, Olkhovik VK, Borchers CH. Isotopically-coded cleavable Crosslinker for studying protein-protein interaction and protein complexes. Molecular & Cellular Proteomics. 2005;**4**:1167-1179

[112] Dokholyan NV. Experimentally-driven protein structure modeling. Journal of Proteomics. 2020;**220**:103777

[113] Brodie NI, Popov KI, Petrotchenko EV, Dokholyan NV, Borchers CH. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. Science Advances. 2017;**3**(7):e1700479. DOI: 10.1126/sciadv.1700479

[114] Serpa JJ, Popov KI, Petrotchenko EV, Dokholyan NV, Borchers CH. Structure of prion β-oligomers as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations. Proteomics. 2019;**21**:e2000298

[115] Brodie NI, Popov KI, Petrotchenko EV, Dokholyan NV, Borchers CH. Conformational ensemble of native α-synuclein in solution as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations. PLoS Computational Biology. 2019;**15**: e1006859

[116] Popov KI, Makepeace KA, Petrotchenko EV, Dokholyan NV, Borchers CH. Insight into the structure of the "unstructured" tau protein. Structure. 2019;**27**:1710-1715.e1714

[117] Marshall JL, Peshkin BN, Yoshino T, Vowinckel J, Danielsen HE, Melino G, et al. The essentials of multiomics. The Oncologist. 2022;**27**:272-284

[118] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology. 2020;**295**:4-15

[119] Bangert P, editor. Machine Learning and Data Science in the Oil and Gas Industry: Best Practices, Tools, and Case Studies. Amsterdam: Elsevier Inc.; 2021

[120] Bangert P. The Necessity for Collaboration Between Data Scientists and Domain Experts the SPE Symposium: Artificial Intelligence - Towards a Resilient and Efficient Energy Industry, virtual. 2021. DOI: 10.2118/ 208634-MS

[121] Bangert P, Moon H, Woo JO, Didari S, Hao H. Active learning performance in labeling radiology images is 90% effective. Frontiers in radiology. 2021

[122] Kleppe A, Skrede OJ, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nature Reviews. Cancer. 2021;**21**:199-211

[123] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence. 2019;**1**:206-215

[124] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy. 2020;**23**:18