

Toward an Integrated Machine Learning Model of a Proteomics Experiment

Benjamin A. Neely,* Viktoria Dorfer,* Lennart Martens,* Isabell Bludau, Robbin Bouwmeester, Sven Degroove, Eric W. Deutsch, Siegfried Gessulat, Lukas Käll, Pawel Palczynski, Samuel H. Payne, Tobias Greisager Rehfeldt, Tobias Schmidt, Veit Schwämmle, Julian Uszkoreit, Juan Antonio Vizcaíno, Mathias Wilhelm, and Magnus Palmblad*



Cite This: <https://doi.org/10.1021/acs.jproteome.2c00711>



Read Online

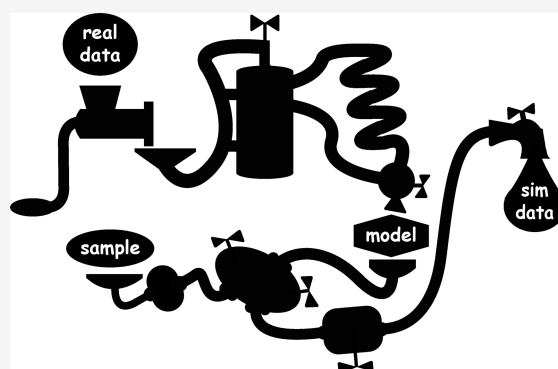
ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: In recent years machine learning has made extensive progress in modeling many aspects of mass spectrometry data. We brought together proteomics data generators, repository managers, and machine learning experts in a workshop with the goals to evaluate and explore machine learning applications for realistic modeling of data from multidimensional mass spectrometry-based proteomics analysis of any sample or organism. Following this sample-to-data roadmap helped identify knowledge gaps and define needs. Being able to generate bespoke and realistic synthetic data has legitimate and important uses in system suitability, method development, and algorithm benchmarking, while also posing critical ethical questions. The interdisciplinary nature of the workshop informed discussions of what is currently possible and future opportunities and challenges. In the following perspective we summarize these discussions in the hope of conveying our excitement about the potential of machine learning in proteomics and to inspire future research.

KEYWORDS: machine learning, deep learning, artificial intelligence, synthetic data, enzymatic digestion, liquid chromatography, ion mobility, tandem mass spectrometry, research integrity



INTRODUCTION

Analytical workflows in proteomics frequently rely on analyzing proteins or peptides by liquid chromatography (LC) and tandem mass spectrometry (MS/MS). Machine learning has been applied to predicting peptide retention times and fragmentation spectra, but recent advances in deep learning have dramatically improved these predictions,^{1–3} as well as modeling other experimental steps, such as enzymatic digestion^{4,5} and ion mobility.⁶ Common to all steps is that the behavior of the proteins, peptides or ions can be predicted from amino acid sequences. However, no model exists that can make realistic predictions of data from a multidimensional separation and analysis of any sample or organism.

Machine learning is set to revolutionize the generation of realistic multidimensional data from arbitrary samples. There are several general application areas of such generative models. For example, when acquiring LC-MS/MS data, acquisition parameters are usually chosen from heuristics and past experience. A machine-learned model predicting LC-MS/MS data for diverse sample types in proteomics could replace guesswork and enable optimal experimental design, including for single cell analysis and other precious biological or clinical

samples. Another application of such a generative model is that synthetic data could provide a benchmark for nearly any sample by creating synthetic data specific to sample processing, multidimensional separation, and data acquisition, regardless of whether the sample had been run before. Currently, system suitability and in-run quality control rely on measuring known samples to evaluate LC-MS/MS performance. If machine learning could be used to predict the expected results of different samples, gradients, and data acquisition, researchers would have a direct measure of the quality of any data set, independent of prior data from experimental reference samples. In turn, this would greatly improve the quality and reusability of nearly any acquired experimental data. Finally, generating an experimentally relevant synthetic ground-truth

Received: November 1, 2022

data set, i.e., known sequences, modifications, or differential abundances, will allow for benchmarking and identifying best performing search algorithms and statistical workflows.

These are three possible applications of such an integrated generative model, but there are countless areas of open scientific questions on future applications of machine learning in proteomics, ranging from data acquisition to biological interpretation. Although machine learning has been applied to individual phases of a proteomics experiment, efforts in combining these into one model have been limited, and there is no comprehensive model that can predict the data for any given sample type, sample preparation, or analysis method required for quality assessment, experimental optimization, or algorithm development. This is unsurprising as no individual researcher or research group has the necessary expertise in all aspects of a proteomics experiment. To foster collaboration between groups and across disciplines, including biology, biochemistry, analytical chemistry, physics, and computer science, we invited experts in these domains to a workshop on proteomics and machine learning held 14 to 18 March, 2022 in Leiden, The Netherlands.

The workshop was designed in early 2021 with an overall goal to discuss and deliver a conceptual design of an integrated machine-learned model of a proteomics experiment, covering all experimental steps from the sampling of the biological system to tandem mass spectrometry (Figure 1). This was accomplished by bringing together researchers who apply machine learning to individual steps of proteomics experiments, such as enzymatic digestion, chromatography, ion mobility, and tandem mass spectrometry, with experts on data repositories and open data formats, as well as practitioners of proteomics representing the end users of machine learned models. The general areas defined above provided a concrete framework for the discussions, including potential misuse of such models, i.e., in data fabrication, and possible social and engineering solutions to mitigate these risks. This summary of our workshop is not meant to be an exhaustive review of machine learning in proteomics, and specialties such as spatial proteomics or mass spectrometry imaging are not covered. We begin our summary of the workshop discussions by looking at trending topics in machine learning generally, and then ask how the proteomes themselves can be predicted.

TRENDING TOPICS IN MACHINE LEARNING

Machine learning, especially deep learning,^{7,8} has made remarkable progress in the past decade. Classically, machine learning could be roughly divided into supervised and unsupervised approaches, while current deep learning advancements have blurred these lines considerably and instead are easier to understand by delineating by use-case (i.e., generating data) and underlying model architecture (i.e., convolutional neural network). During the workshop, discussions were focused on current and future possibilities related to proteomic applications. This section consequently details some of the current machine learning trends that have the greatest potential for impact in proteomics.

Perhaps the most well-known machine learning application in modern day proteomics involves dimensionality reduction techniques such as t-SNE⁹ and UMAP,¹⁰ akin to classically defined unsupervised learning. Both approaches rely on an autoencoder architecture to reduce data noise and ambiguity of high-dimensional data. Notable uses of these approaches includes exploring relevant clusters in single cell proteomics

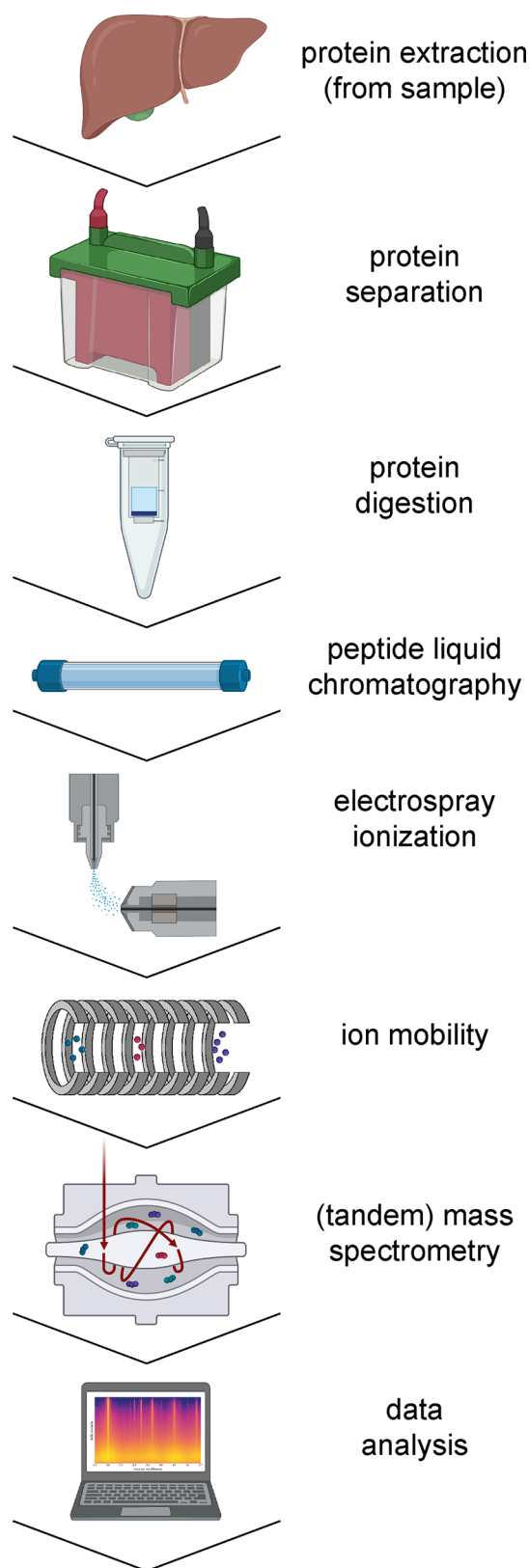


Figure 1. Some common steps in proteomics workflows corresponding to the workshop discussion topics and sections herein. Some icons made using BioRender.com.

data as well as inferring protein–protein interaction networks from high-dimensional combined RNA-Seq and proteomics data.¹¹ In addition to dimensionality reduction techniques,

multimodal representation learning¹² is well suited for proteomics due to its ability to integrate features from different modalities and identify attributes that are shared and different. Successful applications, such as Contrastive Language–Image Pretraining (CLIP¹³), can integrate image and text data into a shared embedding by using a contrastive loss function that trains a model with pairs of positive and negative examples from the two modalities (image and text). Specific applications to mass spectrometry data include SpeCollate,¹⁴ GLEAMS,¹⁵ and yHydra,¹⁶ which are models that learn to jointly embed spectra and peptide sequence information into a shared latent representation. These techniques are pushing the boundaries of what is possible in data prediction and analysis in emerging proteomic techniques.

In addition to interrogating existing data, maybe the most interesting deep machine learning applications, and much of the focus of this paper, involve data generation. Generative approaches, such as generative adversarial networks (GANs¹⁷) and diffusion models,^{18–20} have shown impressive results mainly for image generation but also potential applications in data generation.^{21,22} Broadly, a GAN approach works by having a generator model play a game where it tries to fool its adversary, a classification model that tries to distinguish the fake data by the generator from genuine, real-world data. But training GAN networks are notoriously difficult because the competition between the two models can become unstable and training stops before reaching a satisfying optimum. Recently, GANs have been outpaced by a new class of generative models called diffusion or score-based models¹⁸ that learn a denoising function by applying it multiple times to a signal that is distorted on purpose, similar to denoising autoencoders. With ongoing optimization of speed and more resources, it is likely diffusion-based models will become ubiquitous for countless applications, including proteomics, in the coming years.

Until recently, each data type had a specific neural network layer architecture that it worked best with [i.e., convolutional neural networks (CNNs) were used mostly for images, while recurrent neural networks (RNNs) were used mostly for natural language text sequences], but a new universal layer architecture, the transformer, can work with any input and output data modality. Briefly, this is accomplished by using an attention-based architecture of stacked dense layers. First introduced in 2017,²³ transformers have unified deep learning architectures to become the universal building block for various deep learning approaches. Many modern proteomics models already utilize transformers, including prediction of fragmentation spectra (e.g., Prosit Transformer²⁴) or protein folding (e.g., AlphaFold²⁵), and all of the approaches described above can or already use a transformer. Transformer-based models have captured the public's attention with the ability to generate text (BERT,²⁶ GPT-3²⁷) or images (DALL-E-2²⁸ and DreamStudio/Stable Diffusion¹⁹) from text prompts, and it is exciting to think of similar capabilities with respect to proteomic data generation. In addition to replicating sequential processes in a generative manner, transformers have also been used to successfully predict steps leading to an end point. Retrosynthesis²⁹ is a relevant example, which seeks to determine reaction steps to reach a desired compound. Notably, state-of-the-art models often almost exclusively use transformers and do not rely on other concepts such as convolutional or recurrent layers.

While deep learning is currently the dominant technology in machine learning, it may be excessive for certain prediction

tasks. There is a tendency to value solutions based on these complex algorithms currently in vogue more than simpler models that perform equally well, even if the simplest solution is recommended.³⁰ There is likewise an issue of marginal gains when making very small, incremental improvements in prediction accuracy and/or sensitivity, without being able to show that these improvements have actual value in real life application, demonstrated by improved generalization and improved end result. The [Benchmark Data Sets for Machine Learning in Proteomics](#) section below describes tools and data sets that may help evaluate machine learning models. Finally, as these trending techniques are now applied in proteomics, new methods for machine learning are developed so rapidly that even the most visionary outlook on such topics is likely to quickly be outpaced.

■ PROTEOMES AND HOW TO PREDICT THEM

The ability to predict data from biology, and vice versa, is a primary focus of the topics discussed herein. Specifically, the ability to define the proteome of any species, sample type, and health status *a priori* is necessary for any downstream data prediction steps. Therefore, it is essential to reiterate concepts and highlight current resources of annotated genomes and tissue/biofluid/cell-specific proteome predictions (or empirical data). A proteome is broadly defined as the identity and abundance of proteoforms, which includes isoforms and post-translationally modified forms of each protein, in a given sample. The simplest representation of the protein sequences comprising the unmodified proteome may be derived from an annotated genome of a species, in humans encompassing approximately 20 000 canonical sequences.^{31,32} Recently, there is a greater appreciation for individual variation within a species, emphasizing the need for pan-genomes or study-specific genomes to accurately define the sequence space of the proteome. Beyond humans and common model systems such as mouse, zebrafish, and *Arabidopsis*, there is an explosion of genome annotations available from NCBI RefSeq,³³ UniProtKB,³⁴ and Ensembl,³⁵ and third-party entities like DNA Zoo³⁶ and the EarthBioGenome Project.³⁷ Together, these resources provide the requisite search space for proteomic analysis in nearly any species or environmental sample.

Beyond the catalog of possible proteins in a species' proteome, each tissue, biofluid, and cell-type contains a unique repertoire of proteins at varying abundance,^{38,39} which is also affected by health status (or organism level phenotype). Even so-called “housekeeping proteins” that are detectable in most cell types may vary in abundance.³⁸ While there are various cell atlases and biofluid projects based on experimental proteomic data (Expression Atlas,^{40,41} Genotype-Tissue Expression (GTEx) proteome map,⁴² Human Plasma Proteome,⁴³ Human Protein Atlas,⁴⁴ Human Secretome,⁴⁵ Immunological Proteome Resource,⁴⁶ PeptideAtlas,⁴⁷ and ProteomicsDB⁴⁸), easily accessible lists of tissue/biofluid/cell-specific proteomes with abundance information are rare. Still, it is possible to reprocess proteomic data from public data sets, such as the Wang *et al.* 29 human tissue data set,³⁸ Geiger *et al.* 28 mouse tissue data set,³⁹ or the NCI Proteomics Data Commons cancer-centric data sets,⁴⁹ to derive empirical lists of protein abundances by tissue/biofluid/cell-type and predict what proteins or proteoforms *may* be present in a given sample. Also, it is important to note that there are tissue/cell-specific databases of transcript abundance (such as ARCHS4,⁵⁰ HPA Single Cell Type Atlas,⁵¹ and Tabula Sapiens⁵²), which could

be used as a proxy for protein abundance, albeit with the caveat that transcript and protein abundance is not directly correlated in tissues,⁵³ while in biofluids transcript levels and secreted protein abundance therein are fundamentally disconnected since the source of the secretome is not cells in the biofluid itself.⁴⁵ Although resources like these are largely focused on model organisms, with assumptions, caveats, and caution, tissue/biofluid/cell-type proteomes may be extrapolated between related species. Though all mammals are not primates or rodents, nor all plants are *Arabidopsis* or rice, this can provide an estimation of a proteome when prior knowledge is limited beyond an annotated genome of a related species.

Though these described resources provide confident predictions and thorough evidence of species- and tissue-specific proteomes, the proteomic data itself (i.e., instrumentally measurable signal) will not be a complete measurement of a predicted proteome, since protein extraction, protein digestion, peptide separation techniques, alternatively spliced isoforms, and post-translational modifications (PTMs) will affect the observability of the peptides and the inferred proteoforms. All of these hurdles may be overcome by combining annotated genomes and data repositories with current and future machine learning techniques. This will allow for accurate *in silico* predictions of proteomic data of a given sample type, identification of data from an unknown sample type, and describing experimental and technical artifacts in proteomic data.

A complex remaining task is the prediction of tissue- or organism level changes in protein abundance or post-translational modification, which are intrinsically linked to protein function. Machine learning may complement databases and ontologies to improve functional annotation of proteins, predicting function based on similarity to proteins with known functions in protein families,⁵⁴ inferring function from coregulation of proteins found in large-scale proteomics studies⁵⁵ or integrating protein and RNA-Seq data.¹¹ Machine learning has already been used to predict functional relevance of phosphorylation by combining multiple databases and repositories.⁵⁶ The proteome scale availability of three-dimensional protein structures from AlphaFold enables systematically investigating PTMs in their structural context to further improve our understanding of their functional relevance, and hence our ability to predict their changes.⁵⁷

■ ENZYMATIC DIGESTION

When students first learn about proteases or restriction enzymes, they are invariably told about their specificity as rules or discrete motifs. This is both a necessary and useful simplification. However, in practice, especially in proteomics, the reality is that not all cleavage sites are equal, and that proteolytic digestion is affected by residues outside the canonical motif, as well as PTMs and protein structure. Even if the proteins present in a sample are known, there is still a need to predict the resulting enzymatic peptides.

Predicting cleavage sites of proteins has already a long history, as B. Keil summarized in his analysis of tryptic cleavage in 1992,⁵⁸ stating that trypsin would cleave after arginine (R) or lysine (K) except if it is followed by a proline (P), which is since known as the Keil rule. However, in the last two decades several groups have proven this assumption not always valid, as it has been found that the Keil rule is wrong in about 10% of all cases,⁵⁹ that up to 40% of all tryptic peptides contain missed cleavages,⁶⁰ that cleavage before proline is as common as

cleavage before tryptophan, and even more frequent than cleavage before cysteine.⁶¹ Accurate prediction of cleavage sites has an impact on both types of proteomic analyses, identification, and quantification,⁶² though the impact on protein quantification is much higher. In identification, accurate cleavage prediction can reduce the search space and remove incorrect peptide candidates; however, in quantification, false cleavage estimations can strongly influence the calculated (relative) abundance of inferred proteins.^{63,64}

Recent enzymatic prediction approaches make use of different machine learning techniques, such as random forests^{62,65} and deep learning.^{4,5} In all these approaches, various methods of training data curation are utilized and an amino acid window of size n around a known cleavage site is used as input for the learning algorithm. Most approaches, however, still rely on trypsin as a digestion enzyme, DeepDigest⁵ being one exception. Several other helpful tools have also been developed in the context of enzymatic digestion prediction, one of them being SPACEPro,⁶⁶ a tool that analyzes cleavage efficiency using search results, which could potentially be used to curate training data for cleavage prediction algorithms.

With the recent success of models such as AlphaFold, which has been able to predict, at least to some accuracy, the structures of hundreds of millions of proteins,²⁵ one could imagine including three-dimensional structure and kinetics as input features for machine learning algorithms to predict proteolytic cleavage on a given time scale, down to milliseconds,⁶⁷ in different solvents,⁶⁸ or even predicting structures of fragments and partially denatured proteins. Although attempts have been made in this direction already, there is still a need to develop more models for more enzymes, taking advantage of recent developments and factoring in tertiary structure and PTMs.

■ PROTEIN/PEPTIDE FRACTIONATION

Prediction of how proteins and peptides behave in electrophoretic or chromatographic separations is important in both analytical and preparative contexts, including when analyzing proteomics data from experimental workflows including multiple dimensions of separations or when optimizing preparative methods used to purify recombinant proteins, synthetic peptides, and other products.⁶⁹ Such predictive models can also be used to optimize chromatographic fractionation in proteomics workflows, even if the optimization target is very different than when purifying a single component, for example adjusting chromatographic conditions to distribute the proteins or peptides evenly between fractions with minimal overlap.^{70,71}

Electrophoresis has also been used both preparatively and analytically, and to fractionate proteomes for further analysis by liquid chromatography and mass spectrometry. The predicted electrophoretic migration can be compared to measurements using label-free quantitation, e.g., protein-level SDS-PAGE and peptide-level isoelectric focusing (IEF).⁷² Subsequently, this comparison can flag false positive identifications and covalent complexes (in SDS-PAGE). With increasing speed and sensitivity of mass spectrometers, there is now less perceived need for prefractionation methods in proteomics. However, IEF is still used in a variety of studies where additional separation or targeting of peptides is necessary. It is unknown to what extent machine learning would improve classical predictions, especially since the

isoelectric point (pI) and molecular weight of a protein can be significantly affected by PTMs such as phosphorylation (affecting pI) or glycosylation (affecting both). Unfortunately, current pI calculators are not capable of predicting values for modified peptides. As IEF is still used for separation of modified peptides there is a potential to develop models that can predict the pI of modified peptides. For protein-level IEF under native conditions, a methodology similar to that recently used to calculate protein pK_a values⁷³ using AlphaFold may be adopted.

In any continuous separation method, predicting distributions of analytes in discrete collected fractions is trivial given a prediction of the continuous retention or migration time, though special attention to peak widths and tailing may be warranted to accurately capture the overlap between fractions. A special case of protein and peptide separation methods is the binary fractionation or enrichment of a subproteome or modification of interest, e.g., phosphorylation using titanium dioxide, immobilized metal affinity chromatography, affinity chromatography, or strong cation exchange, or glycosylation using hydrophilic interaction chromatography. Modern machine learning methods have not been applied to predict protein or peptide distribution in such fractionation approaches. Regardless, the direct benefits of accurately modeling protein and peptide behavior in chromatographic or electrophoretic fractionation methods warrants a need for such models, however crude, in our toolbox to enable data simulation of a wide range of proteomics experiments.

■ LIQUID CHROMATOGRAPHY

Whereas the above dimensions of protein and peptide separation are increasingly used in special cases, such as enriching a part of the proteome, reversed-phase liquid chromatography is ubiquitously hyphenated with mass spectrometry in proteomics. Most commonly, the liquid chromatograph is physically coupled inline and online with the mass spectrometer through an electrospray interface, although fraction collection and off-line mass spectrometry via electrospray or MALDI is also possible. Online hyphenations treat time as a continuous variable, whereas in off-line methods time is discretized in intervals and typically not used in the further analyses. For a review on models and molecular simulation studies of reversed-phase liquid chromatography, see Lindsey et al.⁷⁴

Models for retention time predictions have become increasingly complex since the seminal work by Meek published in 1980,⁷⁵ driven by increasing availability and quality of training data with the advent of mass spectrometry-based proteomics. Peptide retention time predictions were first used to assist mass-spectrometry based identification in 2002.⁷⁶ Artificial neural networks trained on amino acid compositions were introduced in 2003⁷⁷ and later extended to take the actual sequence into account.⁷⁸ Other contemporary models used sequence-derived features rather than raw sequences as input, such as SSRCalc⁷⁹ and ELUDE.⁸⁰ More recently, deep neural networks have been trained on large, high-quality, data sets, resulting in even more accurate predictions.^{2,81} Recent implementations of these models have even been shown to even be able to predict previously unseen modifications.^{3,82}

None of these methods attempt to predict elution profiles of peptides, though Afkham et al.⁸³ compared experimental elution profiles with estimated uncertainties in retention time

prediction in their GPTIME model. Accurate modeling of elution profiles is critical in generating realistic synthetic data for optimizing proteomics experiments. Examples of the characteristics that should be modeled are the slope and length of the programmed gradient, the length of dynamic exclusion windows in data-dependent acquisition, and the window size/scan speed in data-independent experiments. Elution simulation here complements actual experiments and, when trained on data on the same chromatographic system and column, could be expected to produce very accurate predictions of retention times as well as chromatographic peak shapes. In targeted proteomics, elution profile prediction would be useful for minimizing interference when selecting peptides and transitions, as well as assist extracted ion chromatogram peak integration for robust quantitation. Elution profile predictions are also required to accurately model chimeric spectra (spectra containing product ions from multiple peptides) in synthetic LC-MS/MS data. These predicted chimeric spectra are useful for scoring identifications in the spectrum. Furthermore, data independent acquisition (DIA) specifically greatly benefits from knowing exactly what peptides contributed to a fragmentation spectrum.^{84,85}

Chromatograms can be aligned using shared features, i.e., common peptides.^{86,87} Internal retention time standards⁸⁸ have become popular in recent years. Although these add little information on the chromatographic separation to the thousands of peptides already in the sample, they simplify comparing data sets and automating data processing by allowing extrapolation to an indexed retention time (iRT) space. Accurate retention time and elution profile models can not only replace these internal standards for chromatographic alignment and system checks, but if the models are interpretable, they can also assist in troubleshooting, e.g., suggest if the mobile phase pH,⁸⁹ temperature, or flow rate is wrong. Whether the goal is to optimize the chromatographic separation or other steps of a proteomics experiment, aligning chromatograms, or rescoring peptide-spectrum matches, realistic LC-MS/MS data simulation will benefit from realistic modeling of the chromatographic behavior of peptides, including elution profiles, and dependence on mobile phase and gradients. This also applies to modified peptides and other modes of chromatography that can be interfaced with mass spectrometry.

■ ION MOBILITY

Ion mobility is increasingly being used in proteomics as a fast separation or trapping method between the ion source and mass analyzer. Most, if not all, major mass spectrometry vendors now integrate ion mobility in their high-end instruments, although the technical implementations work along different principles (e.g., FAIMS, SLIM, TIMS, and TWIMS).⁹⁰ The resolution of the different methods varies substantially,⁹¹ and higher resolution measurements are likely to require more complex algorithms to fit the data. These high-resolution ion mobility measurements are able to differentiate between isomeric structures⁹² that are unlikely to be correctly predicted by simpler models. This means that each ion mobility technique may require a different model, where simplicity while modeling most of the information should be preferred.

If used as a separation device, the ion mobility or arrival times can be calibrated into collisional cross sections (CCSs). Machine learning has already been used to predict CCS values

for different classes of analytes^{6,93,94} and integrated in software for identifying unknown compounds.⁹² As both experimental resolution of ion mobility devices and accuracy of machine-learned models increase, the value of CCS prediction also increases. In cross-linking experiments, CCS values can be used to distinguish cross-linked peptides from unlinked peptides of similar mass-to-charge ratio.⁹⁵ Comparing measured and predicted CCS values has also been used to interrogate protein structure and dynamics, including protein complexes.^{96,97}

For some classes of molecules, CCS is highly correlated with the mass-to-charge ratio, limiting the usefulness of CCS prediction for the identification of unknowns or distinguishing between closely related species. Such lack of orthogonality should always be considered when evaluating the added value of applying machine learning in data analysis workflows. However, to generate realistic synthetic data from experiments including ion mobility separations requires at least some model of analyte behavior.

Ion mobility is becoming increasingly popular in bottom-up proteomics specifically. While the benefits to data acquisition are clear and significant by generating cleaner and more interpretable tandem mass spectra, the impact of predictions of ion mobility behavior or CCS values for peptide identification in untargeted experiments is currently limited, even if machine learned models are quite accurate.⁶ In large search spaces, such as variable PTMs, open modifications, or metaproteomics, CCS predictions could be valuable additions in analysis pipelines to reduce the number of possible candidates for peptide-spectrum matching. For the application of predicted CCS values in open searches, models should be developed that can accurately predict CCS for modified peptides.

■ MASS SPECTROMETRY

There are many choices in mass spectrometry-based proteomics. Top-down or bottom-up? With MS1 or MS2? Using CID or ETD? Low- or high-energy? Data-dependent or data-independent acquisition? Each combination is a different type of experiment, generating different information and requiring a different model to be trained.

The first stage of mass spectrometry, MS1, is acquired to provide accurate mass measurement of intact peptide ions, trigger data-dependent events, and generate quantitative information in experiments such as SILAC. Monoisotopic masses of intact peptide ions and the fragments thereof are easily calculated with sufficient precision based on atomic mass. However, monoisotopic mass is one of few properties of mass spectra that can be so easily calculated. Other isotopic peaks have contributions from multiple isotopologues, the relative abundance of which depend on sample origin, e.g., the fraction ¹³C in plants varies measurably with photosynthetic pathway.⁹⁸ However, for most intents and purposes, isotopic distributions can be calculated from the elemental composition of the peptides and convolved by a theoretical or experimentally sampled peak shape that only depends on the mass-to-charge ratio.

The second stage, tandem mass spectrometry or MS2, is considerably more challenging to predict, and has been the topic of intense research going back at least to 1964 with the DENDRAL software for reconstructing molecular structures from fragment spectra.⁹⁹ This was one of the first “expert systems” and therefore has a special place in the history of artificial intelligence. Tryptic peptide fragmentation patterns

have been gradually refined from uniform predicted intensities for all b- and y-ions¹⁰⁰ to separate intensities for b- and y-ions to intensities dependent on the neighboring residues. More recently, machine learning models such as MS²PIP,¹ Prosit² and others^{101,102} have been shown to produce even more accurate fragmentation pattern predictions. After calibrating the actual collision energy (versus instrument readback), Prosit is able to predict the intensities of b- and y-ions very close to experimental data.

Prediction of tandem mass spectra has many applications in proteomics. Accurate intensity predictions can replace flat intensities or simple models in peptide-spectrum matching algorithms. It is also possible to predict the peptide sequence directly from the spectra, rather than the other way around (i.e., *de novo* sequencing). Models that take collision energy into account can be used to optimize *in silico* collision energies for every targeted peptide in selected-reaction monitoring (SRM; and MRM/PRM), something which is extremely laborious to do experimentally. Simulation of tandem mass spectrometry of peptides is also a key component when simulating realistic proteomics data. This requires modeling the variability in fragmentation, which varies from peptide to peptide and depends on collision energies. In the future, models incorporating PTM prediction, coisolation, nontryptic, and semitryptic digestion, *etc.* will help identify possible peptides to clearly define and optimize model output.

■ PEPTIDE OBSERVABILITY AND PROTEOTYPICITY

When combined, models for all experimental steps described above may be used to predict whether a peptide in a given sample will be detected by the mass spectrometer and “observed”. Different but related concepts and terms exist in this context, e.g., peptide observability or detectability, peptide quantifiability, or proteotypicity. There are different interpretations of these concepts, but key terms are defined as used in the context of this paper below. Peptide observability is the probability that a certain peptide can be identified in a certain sample given that the protein is present (at some level). Often, this probability is seen as binary classification, with the two classes of peptides termed “flyers” and “non-flyers”.¹⁰³ Contrastingly, proteotypicity can be defined as the number of samples with a certain peptide divided by the number of samples with the protein containing that peptide.²

A peptide from a protein present in the biological sample may not be observable at all, for reasons such as poor protein extraction and solubilization, digestion efficiency, PTMs, peptide or protein degradation, suppression by coeluting peptides, stochastic sampling for fragmentation, and biases in the search algorithm. Even when a peptide is detected in a sample, it does not necessarily mean it is quantifiable.¹⁰⁴

Modeling peptide observability requires modeling all the steps from the sample to the mass spectrometer, including peptide or protein fractionation, and enzymatic digestion. However, most proteomics data in repositories either lack these dimensions or the machine-readable metadata to use this information for machine learning, such as standardized and complete sample preparation protocols or the conditions and duration of proteolytic digestion. In general, a prediction model for peptide observability should be able to distinguish between a biological (i.e., nonrandom) missingness of a peptide and a technical (i.e., random) missingness. Proteins that are commonly seen in experiments where protein properties and sample preparation protocols are well-known

and which yield an appropriate peptide distribution could be used as conditional input for such a peptide observability model. Ideal data sets for training models of peptide observability would be instrument-specific repetitive acquisitions from different institutes, different tissues, and different preprocessing methods, e.g., duration of proteolytic digestion. In addition, data sets of synthetic proteins including proteoforms and real-negative samples are helpful.

Several research groups have already addressed the problem of modeling peptide observability.^{103,105–107} Pino and co-workers created a model that combines observability with peptide ionization properties.¹⁰⁴ In 2016, Edfors et al. determined gene-specific coefficients correlating mRNA and protein levels across 20 human tissues and cell lines as measured by RNA-Seq and SRM.⁵³ Recently, Dincer et al. used a deep neural network model, Pepper, to derive sequence-specific coefficients describing the quantitative relationship (or bias) between the observed and measured peptide abundance, finding that the adjusted measurements correlate better than the unadjusted measurements with the RNA-Seq data in quantitative mass spectrometry data.¹⁰⁸

To the best of our knowledge, no models currently predict the proteotypicity in samples enriched for subsets of peptides, e.g., phosphopeptides. Such prediction is challenging, in part due to the large experimental variability of the enrichment step. Prior knowledge of verified modification sites could be helpful in this case.

However, several applications can already take advantage of peptide observability prediction models. In intensity-based quantitation such as iBAQ,¹⁰⁹ the number of observable peptides of a protein can be used to adjust the derived protein abundance. These methods would directly benefit from better estimates on peptide observability. The absence of peptides with high observability is more significant than the absence of those with low observability and may warrant further investigation. Conversely, and although unlikely to be identified as such, peptides with low predicted observability are likely problematic as candidate biomarkers, and high or consistent observability could be used to prioritize candidates.

In peptide identification by spectral libraries and database search engines, information on observability can be used as prior probabilities in peptide-spectrum matching, or to reduce the search space to peptides that are observable under the conditions of the experiment. During protein inference, proteins that cannot be unambiguously identified due to low observability of unique peptides can be excluded.

Although some work has already been performed in this context, there is still much work to be done. As always in machine learning, a variety of training data will be necessary to generate valid models, accompanied by proper, machine readable metadata, which is equally important. Efforts to collect appropriate training data sets has been ongoing for several decades (see [Benchmark Data Sets for Machine Learning in Proteomics](#)), and finally first steps have been taken to enhance the availability of the much-needed corresponding metadata.¹¹⁰

■ MODEL UNCERTAINTY

Typically, machine learning models in proteomics are evaluated on their accuracy rather than the uncertainty, error probability distributions, or confidence intervals that are important when incorporating the models in computational workflows.^{111–114} Software such as Triqler¹¹⁵ and MSE-

QUiP¹¹⁶ model error probability distributions for all steps from peptide-spectrum matching to protein quantification. This alleviates some of the systematic problems with sequential filters and as a consequence improves quantification of low abundance proteins. Other applications for modeling uncertainty can be found in peptide property prediction (e.g., retention time, ion mobility, charge, observability, and fragmentation or spectrum prediction). For retention time prediction, uncertainty estimates can improve the decision of which peptides to include during search.⁸³ Modeling uncertainty of ion intensities is more complex as the combined intensity of one MS2 spectrum represents a joint fragmentation distribution for its isolated precursors. Interestingly, differentiating independent sources of uncertainty can be utilized to improve analyses. For example, there are methods that combine uncertainty estimates to align chromatograms and transfer identifications from one analysis to another.^{117–119} The same principle should be possible to apply to more complex predictions.

Bayesian techniques to model uncertainty in deep neural models include variational inference and Monte Carlo methods. Variational Inference methods learn the posterior distributions over the model's weights, whereas Monte Carlo methods utilize random sampling instead. For example, Monte Carlo dropout¹²⁰ makes several predictions for the same input while randomly ignoring a portion of the model's weights. The resulting predictions are an approximation of the posterior distribution. Repeatedly sampling from different portions of a model is essentially mimicking an ensemble of models with a similar architecture. Such ensemble methods are their own theoretical framework to model uncertainty. Gaussian processes offer a framework for modeling uncertainty without the need for deep architectures. Particularly Gaussian process regression has been successfully used for estimating uncertainty of retention time predictions.⁸³ A more comprehensive overview including other approaches can be found in the review by Abdar et al.¹¹⁴

Knowledge about the uncertainty of predictors could accelerate development of machine learning-driven proteomics. For instance, in training set generation, uncertainty estimates could be utilized to identify which subsets of data need more examples. During targeted proteomics assay development, these measures could be used to improve the exactness of peak prediction and retention time, thereby improving the assays. Confidence could also be transferred from spectral libraries to the peptide-centric analysis of DIA data. In general, computing confidence in single machine learning model predictions of LC-MS/MS peptide behavior, and generation of confidence intervals from peptide spectral matches (re)scoring models (e.g., using SVMs) will benefit protein identification and quantification.

■ SIMULATING LC-MS/MS DATA

Modeling arbitrary proteomics experiments requires models for all steps in [Figure 1](#). Up to this point, the discussion has focused on what is needed to predict the observable peptides from any sample, and the last step is predicting tandem mass spectrometry data. Combined models can produce very realistic simulated data, which has numerous positive use cases, including experimental optimization, quality control using any sample, and benchmarking algorithms and bioinformatic workflows. Generating synthetic LC-MS/MS data in the mzML format is by no means novel. For example,

the OpenMS infrastructure includes MSSimulator¹²¹ for simulating LC-MS/MS data and generating synthetic mzML files from FASTA files. More recently, the SMITER Python library¹²² and Synthedia¹²³ have been used for simulating LC-MS/MS experiments and generate synthetic mzML files. The MaSS-Simulator¹²⁴ and PhosFake (<https://github.com/veitveit/PhosFake>) have also been used to generate synthetic quantitative data or phosphopeptide features, respectively, in other file formats, for the explicit purpose of benchmarking algorithms.

However, recent developments in machine learning, such as GANs and diffusion models, and the specific efforts described above are likely to, when combined, generate far more realistic mzML files. As described herein, by taking into account predicted protein abundances, protein extraction, fractionation, digestion, chromatographic elution profiles of peptides, ionization, ion mobility, precursor selection and chimeric spectra, MS1 and MS2 spectra with all fragment ions, background, systematic and random mass measurement errors, and instrument drift, the synthetic data may be very close to data from real instruments. Such realistic, synthetic data would enable many new applications as discussed, but will also make it far easier to fabricate proteomics data that will be indistinguishable from real data. This inherent risk and possible mitigation were extensively discussed in a dedicated workshop session summarized in the following section.

Existing softwares for synthetic data generation, such as MSSimulator, are highly parametrized and include quite a few instrument aspects, making their realistic settings cumbersome at best. Still, these parameters could be trained using machine learning or dynamic programming methods to produce more realistic tandem mass spectra or any other output from an MS experiment. Software like MSSimulator and SMITER could easily be supplied with state-of-the-art (deep) learning models for predicting tandem mass spectra, e.g., Prosit or MS²PIP, as already done by Synthedia.¹²³

One critical need resulting from this is that simulated LC-MS/MS data in the mzML format should be annotated as such, and the HUPO PSI controlled vocabulary, PSI-MS, extended with suitable terms to describe how the data was simulated, as opposed to generated by a real instrument, or at the very least that the mzML is simulated and the software or model used. Additional considerations for ensuring the integrity and veracity of mzML files are discussed in the section on [Research Integrity](#).

A major new application of synthetic LC-MS/MS data is the *in silico* design and optimization of experiments. Instead of relying on instrument time or precious samples, methods and parameters can be simulated and optimized to yield the maximum amount of information possible, or sufficient information in the shortest possible time to answer a particular research question in an optimization of statistical power. Theoretically, this could reduce development time, improve the offerings in proteomics cores, and increase resulting data richness.

By simulating both samples and experimental configurations, researchers can also estimate what is reasonable data quality for a given sample, system, and method. This is a new dimension to quality control that does not compare data with a previously measured reference or standard but allows quality control metrics to be calculated on a single data set based on what is known about the data set, e.g., the organism, tissue, and experimental method, and comparing the resulting data with

those predicted by machine learning from thousands or even millions of data sets from different organisms, tissues, and experiments. In other words, instead of relying on a specific standard or material to evaluate system suitability or experimental performance, any given sample can provide the same actionable information.

Lastly, realistic, synthetic LC-MS/MS data can be used to test, improve, and benchmark machine learning and classical algorithms and software. This can be accomplished by improving the output from database search engines using simulated spectra, generating ground-truth protein inference data to challenge protein inference algorithms, and generating ground-truth quantitative data to evaluate quantitative proteomics software. Unlike real ground-truth data, synthetic data sets would not have any systematic experimental errors from protein extraction or pipetting and could be created with the click of a button, instead of years of often tedious work preparing samples and curating spectra. Though it may seem that synthetic data may never “look” exactly like real data, it should be noted how advanced deep learning techniques have become at generating text and images that are nearly indistinguishable from those created by humans. We expect that crossing this uncanny valley is closer than it appears. Bridging it will create countless applications, including those listed above, and pose a formidable ethical dilemma.

RESEARCH INTEGRITY

Although the workshop focused on legitimate uses of realistic synthetic data, there is a possibility such data is passed off as real. This topic was discussed in detail, including mitigating social and engineering solutions. As most technologies, machine learning is amoral and can be used for beneficial or nefarious purposes alike. We hope that this pre-emptive discussion helps raise awareness and spur the development of safeguards.

Fraud is a problem of increasing concern in science, as various social and personal pressures entice individuals to make up data and attempt to publish it as legitimate research. In an extensive review of over 20 000 peer-reviewed publications containing images of Western blots, nearly 4% contained inappropriate manipulation.¹²⁵ As this study was limited to Western blots, it only shows the “tip of the iceberg” of data fabrication in the biomedical literature. Unlike images, data fabrication in numerical tables can be subtle and hard to detect. In 2020, Bradshaw and Payne examined methods to detect fabricated numerical omics data,¹²⁶ concluding that methods originally developed to detect fraud in banking and insurance would be applicable to scientific data.

There are already known examples of mass spectrometry data fabrication. In 2020, the University of Liverpool investigated a case of alleged research misconduct, finding falsification of key mass spectrometry data,^{127,128} including in the now-removed Figure 2 in a corrected paper.^{127,129} In this case, the misconduct was confirmed by extrinsic evidence such as equipment logs and financial records revealing the measurements were never performed. It would be naive to think this is an isolated case, given the number of Western blots found to be manipulated, when these were systematically investigated.

The workshop session focused on data fabrication and falsification using generative machine learning.²² Data fabrication could happen at different stages of proteomics experiments: experimental data acquisition, manipulation of

existing raw data, and downstream data analyses. We will not here provide details on how this can be done, but the workshop demonstrated that it is currently possible to fabricate proteomics data in ways that are not all trivial to detect. However, there are also several engineering and social solutions to mitigate risks of realistically simulated data being passed off as real.

Manipulated or fabricated data can be detected by a variety of algorithms leveraging number theory metrics unrelated to study design, such as Benford's Law¹³⁰ used by Bradshaw and Payne,¹²⁶ and Zipf's Law.¹³¹ The workshop demonstrated that some existing models for predicting mass spectra violate some of these laws. Beyond number theory, detection algorithms can be strengthened by domain knowledge and look for expected patterns in different data layers and comparing measured masses, peak shapes and isotopic distributions with the limits imposed by the instrument settings. It is possible to calculate these metrics from public data in repositories to obtain the limits and expected values of these metrics, assuming that this data is in fact real. One can also check for expected correlations in quantitative data, based protein-peptide stoichiometry or coexpression. If data is generated by spiking synthesized peptides of interest into the sample, and then measuring it using an actual mass spectrometer, one may be able to find remnants of the peptide synthesis in the data.

In addition to algorithmic fraud detection, it would be advantageous for the community to support a limited set of tools that generate synthetic data, which are built with data integrity in mind. There are at least two important features of such tools. First, they should target the open standard mzML file format, rather than vendor formats. The latter could be made more immutable by encoding timestamps, location and instrument serial numbers, file hashes, and theoretically even using technologies such as blockchain that prove a particular data set was acquired on a particular instrument in a particular location at a particular time. One drawback is that it then becomes impossible to fully anonymize data. Data anonymization may be preferred when publishing certain raw data, for example from interlaboratory comparisons or ring trials. Fortunately, with many vendors moving into clinical trials, many of their formats are already relatively immutable.

A second safeguard are digital watermarks in synthetic data. This can be accomplished by inserting patterns in the signal noise or features that do not match any known protein sequence (e.g., "FAKEPEPTIDE" amino acid sequence). These watermarks should be cumbersome to remove but not interfere with any data analysis. This safeguard requires that at least some components of the generating software are not provided as open source. As with security features in banknotes, some watermarks can be publicly known and easy to detect, while others would be confidentially shared with journals and repositories, and some known only by the tool developers. Overall, this would allow synthetic data to be used without risking them being mistaken for real.

Beyond engineering solutions, there are also necessary social solutions. Given that data fabrication or duplication can happen at different levels, it is crucial to have access to the data provenance from raw data to reported results, requiring that all data, including raw data in closed vendor formats, and code are made available before publication. Such solutions already exist in the form of the ProteomeXchange repositories¹³² and ability to provide data analysis scripts or notebooks as Supporting Information or on a public server such as OSF (<https://osf.io/>

) or GitHub (<https://github.com>), containing all steps from the raw data in the repository to the figures and tables in the publication. Ensuring this information is actually provided requires social solutions such as journal policies requiring raw data in vendor formats to be deposited and educating reviewers and editors about these topics.

In summary, it is important the proteomics community is aware of the increasing ease with which realistic synthetic proteomics data can be generated using machine learning. The prevalence of problematic Western blots in the literature should be a cause for reflection, as these are far cheaper to perform or repeat than state-of-the-art proteomics experiments requiring million-dollar instruments and reagents for thousands of dollars. At the same time, there are many legitimate use cases for simulating proteomics data, as outlined in the [Introduction](#), and with a combination of engineering and social solutions, the risks that such realistic simulated data is being passed off as real undetected can be reduced.

■ BENCHMARK DATA SETS FOR MACHINE LEARNING IN PROTEOMICS

The workshop discussed the steps needed to develop a unified model to generate synthetic proteomic data, discussed possible applications of this capability, and touched on ethical concerns therein. However, to effectively catalyze this research topic, suitable proteomics data sets must be accessible to machine learning practitioners, as well as machine learning methods to data generators. Accessible and fit-for-purpose benchmark data sets are essential in domains such as machine learning for education and calibration across the field. For instance, the infamous Anderson's *Iris* data set¹³³ continues to be used in most beginner tutorials to demonstrate supervised and unsupervised machine learning based on petal and sepal length and width. More advanced machine learning modelers typically move on to the Titanic data set¹³⁴ with more attributes and larger sample size. These two example data sets are accessible to newcomers and experts alike because of their small size, fit for purpose due to their straightforward attributes, and their frequent use. Together this results in an abundance of available tutorials across different software, languages, and algorithms, all based on the same underlying data set. Having benchmark data sets like these creates a point of entry with a low threshold, easy for beginners and education, but more importantly, these data sets can be used to explore and benchmark current and new machine learning techniques, irrespective of domain knowledge of the data set itself.

Specific to proteomics applications of machine learning, there is a need for similarly accessible and fit for purpose benchmark data sets. Although many machine learning endeavors such as DeepLC³ and Prosit² validate against external data sets from laboratories that were not used during model training, the community has not agreed on specific benchmark data sets. Such benchmark data sets would make comparisons of different models more informative. Many of these deep learning models are also trained on samples from the same data sets, for example the data sets of synthetic peptides from ProteomeTools,¹³⁵ and while these are a great resource for homogenized training, it can be hard to compare across tools without any collectively acknowledged inference source. Additionally, tools have recently been released that make it easier for researchers to extract machine learning-ready data from raw files, such as MS2AI,¹³⁶ but even this is limited

due to the complexity and variety of instruments and software used in proteomics.

In the months following the workshop, participants developed a resource to address this need.¹³⁷ **ProteomicsML.org** provides an online repository of easily ingested data sets with attributes spanning peptide properties and mass spectrometry data types, as well as companion tutorials on training deep learning models. This dynamic resource of proteomics benchmark data sets will be curated similar to the UCI Machine Learning Repository¹³⁸ or MoleculeNet,¹³⁹ and is mirrored on a GitHub repository to enable programmatic access (<https://github.com/ProteomicsML/ProteomicsML>). The benchmark data sets in ProteomicsML.org are optimally sized data matrices, with data import ease and handling as the goal. The specific search settings and filtering used to create each data set is described so that users can also preprocess from source raw mass spectrometry files (as PRIDE¹⁴⁰ or MassIVE¹⁴¹ identifiers), but importantly, this is not required for using the data sets in machine learning applications. As described in prior sections, there are numerous machine learning applications in proteomics, and the data sets at ProteomicsML.org are also organized by application (i.e., retention time, spectrum prediction, ion mobility cross-section, enzymatic digestion, peptide observability/proteotypic peptides, peptide fractionation). The ProteomicsML.org resource will grow with community involvement, including both training and testing data sets by application, and larger data sets to expand alongside computational and instrument advances, helping machine learning experts experiment with proteomics data, and proteomics experts learn about machine learning applications.

CONCLUSIONS

Taken together, the concepts and issues presented here provide a relevant, and hopefully exciting overview of the status of machine learning in proteomics today, and a few possible paths forward. It is clear that machine learning in proteomics has not just exploded in recent years, but is in fact here to stay. The breakthroughs in peptide identification performance alone have been quite impressive, and still new work continues to improve upon the previous state-of-the-art.


Perhaps most notably, the individual models that have been built so far could be assembled into a larger, end-to-end model, which could then predict realistic, synthetic data, which in turn could have many positive uses, as outlined in detail above, but that could also pass for real data if a nefarious person offered such synthetic data up as actual acquired data. Appreciating this capability, albeit theoretical, creates novel issues in proteomics that are deserving of attention. Concrete steps can be taken to safeguard against the potential issues raised by realistic synthetic data, while ensuring that positive uses may be pursued without issue. Rather than wait for the problem to occur (or be noticed), it would be prudent if the field began preemptively setting up appropriate safeguards, especially concerning the low-hanging fruit of adapting standard mass spectrometry file formats to accommodate well-annotated and clearly flagged synthetic data, while also ensuring that instrument raw files are fingerprinted or marked in a way that would allow their provenance to be traced to the actual instrument if needed. The latter comes with its own issues regarding data generator anonymity, and should therefore likely be engineered in such a way as to only be possible upon specific request.


However, before building such comprehensive and believable mass spectrometry data generators, more sophisticated models will likely be necessary. These models in turn will require ever better training data and adequate benchmarking data. Here too, there are several key issues to address moving forward, notably with regard to metadata provisioning. Currently, metadata provisioning in proteomics is scarce at best, and can frequently be found to be incorrect upon *a posteriori* verification.¹⁴² A more diligent annotation of data deposited in the public domain would provide much-needed leverage for a variety of downstream uses, which include building better machine learning models.


The power of machine learning as a nearly universal tool has created a modern-day Maslow's (jack)hammer,¹⁴³ and likewise machine learning is not always the best tool for the job. Many prediction tasks in proteomics do not require machine learning. As a trivial example, theoretical fragment ion masses can be easily calculated by summing the monoisotopic masses of the corresponding amino acid residues and applying the ion-specific mass loss correction. Though machine learning is making bold strides in modeling nearly every aspect of a mass spectrometry-based proteomics experiment from the biological system to interpretation of the results, the gains offered by machine learning should be critically evaluated. The curated data sets of ProteomicsML.org are useful for both beginners and experienced practitioners of machine learning, and for benchmarking and quantifying the gains of new models. With the accelerating growth in size and complexity of proteomics data, machine learning will become increasingly indispensable and fundamentally change the way proteomic data is acquired and interpreted.

AUTHOR INFORMATION

Corresponding Authors

Benjamin A. Neely — National Institute of Standards and Technology, Charleston, South Carolina 29412, United States;  orcid.org/0000-0001-6120-7695; Email: benjamin.neely@nist.gov


Viktoria Dorfer — Bioinformatics Research Group, University of Applied Sciences Upper Austria, 4232 Hagenberg, Austria;  orcid.org/0000-0002-5332-5701; Email: viktoria.dorfer@fh-hagenberg.at

Lennart Martens — VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Health Sciences and Medicine, Ghent University, 9000 Ghent, Belgium;  orcid.org/0000-0003-4277-658X; Email: lennart.martens@ugent.be

Magnus Palmblad — Leiden University Medical Center, 2300 RC Leiden, The Netherlands;  orcid.org/0000-0002-5865-8994; Email: n.m.palmblad@lumc.nl

Authors

Isabell Bludau — Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

Robbin Bouwmeester — VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium; Department of Biomolecular Medicine, Faculty of Health Sciences and Medicine, Ghent University, 9000 Ghent, Belgium;  orcid.org/0000-0001-6807-7029

Sven Degroove — VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium; Department of

Biomolecular Medicine, Faculty of Health Sciences and Medicine, Ghent University, 9000 Ghent, Belgium; orcid.org/0000-0001-8349-3370

Eric W. Deutsch – Institute for Systems Biology, Seattle, Washington 98109, United States; orcid.org/0000-0001-8732-0928

Siegfried Gessulat – MSAID GmbH, 10559 Berlin, Germany

Lukas Käll – Science for Life Laboratory, KTH - Royal Institute of Technology, 171 21 Solna, Sweden; orcid.org/0000-0001-5689-9797

Pawel Palczynski – Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense, Denmark

Samuel H. Payne – Department of Biology, Brigham Young University, Provo, Utah 84602, United States; orcid.org/0000-0002-8351-1994

Tobias Greisager Rehfeldt – Institute for Mathematics and Computer Science, University of Southern Denmark, 5230 Odense, Denmark; orcid.org/0000-0002-1190-9485

Tobias Schmidt – MSAID GmbH, 85748 Garching, Germany; orcid.org/0000-0002-1883-6514

Veit Schwämmle – Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense, Denmark; orcid.org/0000-0002-9708-6722

Julian Uszkoreit – Medical Proteome Analysis, Center for Protein Diagnostics (ProDi), Ruhr University Bochum, 44801 Bochum, Germany; Medizinisches Proteom-Center, Medical Faculty, Ruhr University Bochum, 44801 Bochum, Germany; orcid.org/0000-0001-7522-4007

Juan Antonio Vizcaino – European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom; orcid.org/0000-0002-3905-4335

Mathias Wilhelm – Computational Mass Spectrometry, Technical University of Munich (TUM), 85354 Freising, Germany; orcid.org/0000-0002-9224-3258

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jproteome.2c00711>

Notes

The authors declare the following competing financial interest(s): S.G. and T.S. are co-founders, shareholders and employees of MSAID GmbH, a company that develops software for proteomics. M.W. is founder and shareholder of MSAID GmbH and Omicscouts GmbH, with no operational role in both companies.

ACKNOWLEDGMENTS

The Lorentz Center workshop was funded by the Dutch Research Council (NWO) with generous support from the Leiden University Medical Center, Thermo Fisher Scientific, and the *Journal of Proteome Research* that made this workshop possible. The authors thank Dr. Mike Preuss and Dr. Elisabeth Bik for their inspiring lectures and active participation in discussions on machine learning and research integrity during the workshop. Special thanks go to all the Lorentz Center staff who went above and beyond their duties to make this hybrid workshop a success in COVID times. Identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it

imply that the products identified are necessarily the best available for the purpose.

ABBREVIATIONS

ARCHS4, all RNA-seq and ChIP-seq sample and signature search; BERT, Bidirectional Encoder Representations from Transformers; CCS, collisional cross sections; CID, collision-induced dissociation; CLIP, Contrastive Language–Image Pretraining; DALL-E-2, Machine learned model for generating images from text descriptions (portmanteau of WALL-E and Dali); DENDRAL, Dendritic Algorithm; DIA, data independent acquisition; ELUDE, unknown; wordplay on elution seems likely; Ensembl, just a name; ETD, electron-transfer dissociation; FAIMS, high field asymmetric waveform ion mobility spectrometry; FASTA, derived from the FASTP (P for protein) similarity search, FASTA is short for Fast All, and the file format used became known as fasta.; GANs, generative adversarial networks; GLEAMS, a learned embedding for annotating mass spectra; GPT-3, Generative Pretrained Transformer 3; GPTIME, Gaussian processes time; HPA, Human Protein Atlas; HUPO PSI, Human Proteome Organization Proteomics Standards Initiative; iBAQ, intensity-based absolute quantification; IEF, isoelectric focusing; iRT, indexed retention time; LC, liquid chromatography; MALDI, matrix-assisted laser desorption/ionization; MassIVE, mass spectrometry interactive virtual environment; MaSS-Simulator, made to differentiate from similar tools: MSSimulator and MS-Simulator; MRM, multiple reaction monitoring; mRNA, messenger ribonucleic acid; MS/MS, tandem mass spectrometry; MS2A, mass spec to artificial intelligence; MS²PIP, MS2 peak intensity prediction; MSeQUIP, method for quantifying uncertainty in peptide-spectrum matches; MuSIC, multiscale integrated cell; mzML, official HUPO-PSI standard format for mass spectrometry data, derivative of previous mzXML format, mz (mass-to-charge ratio) and XML (eXtensible Markup Language); NCBI RefSeq, National Center for Biotechnology Information Reference Sequence Database; NCI, National Cancer Institute; OSF, Open Science Framework; pI, negative log (base 10) of isoelectric point; pK_a, negative log (base 10) of acid dissociation constant; PRIDE, Proteomics Identification Database; PRM, parallel reaction monitoring; ProteomicsDB, DB is short for database; PTMs, post-translational modifications; RNA-Seq, ribonucleic acid sequencing; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; SILAC, stable isotope labeling by/with amino acids in cell culture; SLIM, structures for lossless ion manipulation; SMITER, Synthetic mzML writer; SPACE-Pro, Shotgun Proteomic Analysis of Cleavage Efficiency of Proteins; SRM, selected-reaction monitoring; SSRCalc, sequence-specific retention calculator; SUMOylation, small ubiquitin-like modifier modification; SVM, support vector machine; TIMS, trapped ion mobility spectrometry; t-SNE, t-distributed stochastic neighbor embedding; TWIMS, traveling wave ion mobility spectrometry; UCI, University of California, Irvine; UMAP, Uniform Manifold Approximation and Projection; UniProtKB, Universal Protein Resource Knowledgebase

REFERENCES

- (1) Degroove, S.; Maddelein, D.; Martens, L. MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic acids research* **2015**, *43* (W1), W326–30.

- (2) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **2019**, *16* (6), 509–518.
- (3) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroove, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **2021**, *18* (11), 1363–1369.
- (4) Sun, B.; Smialowski, P.; Straub, T.; Imhof, A. Investigation and Highly Accurate Prediction of Missed Tryptic Cleavages by Deep Learning. *J. Proteome Res.* **2021**, *20* (7), 3749–3757.
- (5) Yang, J.; Gao, Z.; Ren, X.; Sheng, J.; Xu, P.; Chang, C.; Fu, Y. DeepDigest: Prediction of Protein Proteolytic Digestion with Deep Learning. *Anal. Chem.* **2021**, *93* (15), 6094–6103.
- (6) Meier, F.; Köhler, N. D.; Brunner, A.-D.; Wanka, J.-M. H.; Voytik, E.; Strauss, M. T.; Theis, F. J.; Mann, M. Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.* **2021**, *12* (1), 1185.
- (7) Bengio, Y.; Courville, A.; Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* **2013**, *35* (8), 1798–1828.
- (8) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436–44.
- (9) Maaten, L. v. d.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learning Res.* **2008**, *9* (86), 2579–2605.
- (10) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, DOI: 10.48550/arXiv.1802.03426.
- (11) Koutrouli, M.; Lindez, P. P.; Bouwmeester, R.; Martens, L.; Jensen, L. J.; FAVA: High-quality functional association networks inferred from scRNA-seq and proteomics data. *bioRxiv* **2022**, DOI: 10.1101/2022.07.06.499022.
- (12) Li, Y.; Yang, M.; Zhang, Z. A Survey of Multi-View Representation Learning. *IEEE Transactions on Knowledge and Data Engineering* **2019**, *31* (10), 1863–1883.
- (13) Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*; PMLR: 2021; pp 8748–8763.
- (14) Tariq, M. U.; Saeed, F. SpeCollate: Deep cross-modal similarity network for mass spectrometry data based peptide deductions. *PLoS One* **2021**, *16* (10), e0259349.
- (15) Bittremieux, W.; May, D. H.; Bilmes, J.; Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. *Nat. Methods* **2022**, *19* (6), 675–678.
- (16) Altenburg, T.; Muth, T.; Renard, B. Y. yHydra: Deep Learning enables an Ultra Fast Open Search by Jointly Embedding MS/MS Spectra and Peptides of Mass Spectrometry-based Proteomics. *bioRxiv* **2021**, DOI: 10.1101/2021.12.01.470818.
- (17) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* **2020**, *63* (11), 139–144.
- (18) Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, DOI: 10.48550/arXiv.2010.02502.
- (19) Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022; pp 10684–10695.
- (20) Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *CoRR* **2020**, DOI: 10.48550/ARXIV.2006.11239.
- (21) Yelmen, B.; Decelle, A.; Ongaro, L.; Marnetto, D.; Tallec, C.; Montinaro, F.; Furtlehner, C.; Pagani, L.; Jay, F. Creating artificial human genomes using generative neural networks. *PLoS Genet* **2021**, *17* (2), e1009303.
- (22) Wang, Z.; Wang, J.; Kourakos, M.; Hoang, N.; Lee, H. H.; Mathieson, I.; Mathieson, S. Automatic inference of demographic parameters using generative adversarial networks. *Mol. Ecol Resour* **2021**, *21* (8), 2689–2705.
- (23) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**, NIPS: 2017.
- (24) Ekvall, M.; Truong, P.; Gabriel, W.; Wilhelm, M.; Käll, L. Prosit Transformer: A transformer for Prediction of MS2 Spectrum Intensities. *J. Proteome Res.* **2022**, *21* (5), 1359–1364.
- (25) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589.
- (26) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, DOI: 10.48550/arXiv.1810.04805.
- (27) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Advances in Neural Information Processing Systems* **2020**, NeurIPS: 2020; pp 1877–1901.
- (28) Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, DOI: 10.48550/arXiv.2204.06125.
- (29) Karpov, P.; Godin, G.; Tetko, I. V. In *A Transformer Model for Retrosynthesis*; Springer International Publishing: Cham, 2019; pp 817–830.
- (30) Vleugels, R. P. M Search and Chemical Ontology Plotting Environment. <https://github.com/ReinV/SCOPE>.
- (31) Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic acids research* **2017**, *45* (D1), D177–d182.
- (32) Omenn, G. S.; Lane, L.; Overall, C. M.; Paik, Y.-K.; Cristea, I. M.; Corrales, F. J.; Lindskog, C.; Weintraub, S.; Roehrl, M. H. A.; Liu, S.; Bandeira, N.; Srivastava, S.; Chen, Y.-J.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2021**, *20* (12), S227–S240.
- (33) Pruitt, K.; Brown, G.; Tatusova, T.; Maglott, D. The Reference Sequence (RefSeq) Database. In *NCBI Handbook [Internet]*; McEntyre, J., Ostell, J., Eds.; National Center for Biotechnology Information (US): Bethesda (MD), 2002.
- (34) The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489.
- (35) Cunningham, F.; Allen, J. E.; Allen, J.; Alvarez-Jarreta, J.; Amode, M. R.; Armean, I. M.; Austine-Orimoloye, O.; Azov, A. G.; Barnes, I.; Bennett, R.; Berry, A.; Bhai, J.; Bignell, A.; Billis, K.; Boddu, S.; Brooks, L.; Charkhchi, M.; Cummins, C.; Da Rin Fioretto, L.; Davidson, C.; Dodiya, K.; Donaldson, S.; El Houdaigui, B.; El Naboulsi, T.; Fatima, R.; Giron, C. G.; Genez, T.; Martinez, J. G.; Guirarro-Clarke, C.; Gymer, A.; Hardy, M.; Hollis, Z.; Hourlier, T.; Hunt, T.; Juettemann, T.; Kaikala, V.; Kay, M.; Lavidas, I.; Le, T.; Lemos, D.; Marugán, J. C.; Mohanan, S.; Mushtaq, A.; Naven, M.; Ogeh, D. N.; Parker, A.; Parton, A.; Perry, M.; Piližota, I.; Prosovetskaia, I.; Sakthivel, M. P.; Salam, A. I. A.; Schmitt, B. M.; Schuilenburg, H.; Sheppard, D.; Pérez-Silva, J. G.; Stark, W.; Steed, E.; Sutinen, K.; Sukumaran, R.; Sumathipala, D.; Suner, M.-M.; Szpak, M.; Thormann, A.; Tricomi, F. F.; Urbina-Gómez, D.; Veidenberg, A.; Walsh, T. A.; Walts, B.; Willhoft, N.; Winterbottom, A.; Wass, E.; Chakiachvili, M.; Flint, B.; Frankish, A.; Giorgetti, S.; Haggerty, L.; Hunt, S. E.; Iisley, G. R.; Loveland, J. E.; Martin, F. J.; Moore, B.; Mudge, J. M.; Muffato, M.; Perry, E.; Ruffier, M.; Tate, J.; Thybert,

- D.; Trevanion, S. J.; Dyer, S.; Harrison, P. W.; Howe, K. L.; Yates, A. D.; Zerbino, D. R.; Flicek, P. Ensembl 2022. *Nucleic acids research* **2022**, 50 (D1), D988–D995.
- (36) Dudchenko, O.; Shamim, M. S.; Batra, S. S.; Durand, N. C.; Musial, N. T.; Mostofa, R.; Pham, M.; St; Hilaire, B. G.; Yao, W.; Stamenova, E. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. *bioRxiv* **2018**, 254797.
- (37) Lewin, H. A.; Robinson, G. E.; Kress, W. J.; Baker, W. J.; Coddington, J.; Crandall, K. A.; Durbin, R.; Edwards, S. V.; Forest, F.; Gilbert, M. T. P.; Goldstein, M. M.; Grigoriev, I. V.; Hackett, K. J.; Haussler, D.; Jarvis, E. D.; Johnson, W. E.; Patrinos, A.; Richards, S.; Castilla-Rubio, J. C.; van Sluys, M.-A.; Soltis, P. S.; Xu, X.; Yang, H.; Zhang, G. Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, 115 (17), 4325–4333.
- (38) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.-H.; Meng, C.; Frejno, M.; Schmidt, T.; Schnatbaum, K.; Wilhelm, M.; Ponten, F.; Uhlen, M.; Gagneur, J.; Hahne, H.; Kuster, B. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **2019**, 15 (2), e8503.
- (39) Geiger, T.; Velic, A.; Macek, B.; Lundberg, E.; Kampf, C.; Nagaraj, N.; Uhlen, M.; Cox, J.; Mann, M. Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol. Cell Proteomics* **2013**, 12 (6), 1709–1722.
- (40) Papatheodorou, I.; Moreno, P.; Manning, J.; Fuentes, A. M.-P.; George, N.; Fexova, S.; Fonseca, N. A.; Füllgrabe, A.; Green, M.; Huang, N.; Huerta, L.; Iqbal, H.; Jianu, M.; Mohammed, S.; Zhao, L.; Jarnuczak, A. F.; Jupp, S.; Marioni, J.; Meyer, K.; Petryszak, R.; Prada Medina, C. A.; Talavera-López, C.; Teichmann, S.; Vizcaino, J. A.; Brazma, A. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **2019**, 48 (D1), D77–D83.
- (41) Prakash, A.; García-Seisdedos, D.; Wang, S.; Kundu, D. J.; Collins, A.; George, N.; Moreno, P.; Papatheodorou, I.; Jones, A. R.; Vizcaino, J. A. An integrated view of baseline protein expression in human tissues. *J. Proteome Res.* **2022**, (Article ASAP). DOI: 10.1021/acs.jproteome.2c00406.
- (42) Jiang, L.; Wang, M.; Lin, S.; Jian, R.; Li, X.; Chan, J.; Dong, G.; Fang, H.; Robinson, A. E.; Consortium, G.; Snyder, M. P. A Quantitative Proteome Map of the Human Body. *Cell* **2020**, 183 (1), 269–283.e19.
- (43) Schwenk, J. M.; Omenn, G. S.; Sun, Z.; Campbell, D. S.; Baker, M. S.; Overall, C. M.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J. Proteome Res.* **2017**, 16 (12), 4299–4310.
- (44) Uhlen, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szegedy, C. A.-K.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P.-H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Pontén, F. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, 347 (6220), 1260419.
- (45) Uhlen, M.; Karlsson, M. J.; Hober, A.; Svensson, A.-S.; Scheffel, J.; Kotol, D.; Zhong, W.; Tebani, A.; Strandberg, L.; Edfors, F.; Sjöstedt, E.; Mulder, J.; Mardinoglu, A.; Berling, A.; Ekblad, S.; Dannemeyer, M.; Kanje, S.; Rockberg, J.; Lundqvist, M.; Malm, M.; Volk, A.-L.; Nilsson, P.; Månberg, A.; Dodig-Crnkovic, T.; Pin, E.; Zwahlen, M.; Oksvold, P.; von Feilitzen, K.; Häussler, R. S.; Hong, M.-G.; Lindskog, C.; Ponten, F.; Katona, B.; Vu, J.; Lindström, E.; Nielsen, J.; Robinson, J.; Ayoglu, B.; Mahdessian, D.; Sullivan, D.; Thul, P.; Danielsson, F.; Stadler, C.; Lundberg, E.; Bergström, G.; Gummesson, A.; Voldborg, B. G.; Tegel, H.; Hober, S.; Forsström, B.; Schwenk, J. M.; Fagerberg, L.; Sivertsson, Å. The human secretome. *Sci. Signal* **2019**, 12 (609), eaaz0274.
- (46) Brenes, A. J.; Hukelmann, J. L.; Spinelli, L.; Howden, A. J.; Marchingo, J. M.; Sinclair, L. V.; Rollings, C.; James, O. J.; Phair, I. R.; Matthews, S. P. The Immunological Proteome Resource. *bioRxiv* **2022**, DOI: 10.1101/2022.08.29.505666.
- (47) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, 34 (Database issue), D655–D658.
- (48) Schmidt, T.; Samaras, P.; Frejno, M.; Gessulat, S.; Barnert, M.; Kienegger, H.; Krcmar, H.; Schlegel, J.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M. ProteomicsDB. *Nucleic acids research* **2018**, 46 (D1), D1271–D1281.
- (49) Heath, A. P.; Ferretti, V.; Agrawal, S.; An, M.; Angelakos, J. C.; Arya, R.; Bajari, R.; Baqar, B.; Barnowski, J. H. B.; Burt, J.; Catton, A.; Chan, B. F.; Chu, F.; Cullion, K.; Davidsen, T.; Do, P.-M.; Dompierre, C.; Ferguson, M. L.; Fitzsimons, M. S.; Ford, M.; Fukuma, M.; Gaheen, S.; Ganji, G. L.; Garcia, T. I.; George, S. S.; Gerhard, D. S.; Gerthoffert, F.; Gomez, F.; Han, K.; Hernandez, K. M.; Issac, B.; Jackson, R.; Jensen, M. A.; Joshi, S.; Kadam, A.; Khurana, A.; Kim, K. M. J.; Kraft, V. E.; Li, S.; Lichtenberg, T. M.; Lodato, J.; Lolla, L.; Martinov, P.; Mazzone, J. A.; Miller, D. P.; Miller, I.; Miller, J. S.; Miyauchi, K.; Murphy, M. W.; Nullet, T.; Ogawa, R. O.; Ortuño, F. M.; Pedrosa, J.; Pham, P. L.; Popov, M. Y.; Porter, J. J.; Powell, R.; Rademacher, K.; Reid, C. P.; Rich, S.; Rogel, B.; Sahni, H.; Savage, J. H.; Schmitt, K. A.; Simmons, T. J.; Sislow, J.; Spring, J.; Stein, L.; Sullivan, S.; Tang, Y.; Thiagarajan, M.; Troyer, H. D.; Wang, C.; Wang, Z.; West, B. L.; Wilmer, A.; Wilson, S.; Wu, K.; Wysocki, W. P.; Xiang, L.; Yamada, J. T.; Yang, L.; Yu, C.; Yung, C. K.; Zenklusen, J. C.; Zhang, J.; Zhang, Z.; Zhao, Y.; Zubair, A.; Staudt, L. M.; Grossman, R. L. The NCI Genomic Data Commons. *Nat. Genet.* **2021**, 53 (3), 257–262.
- (50) Lachmann, A.; Torre, D.; Keenan, A. B.; Jagodnik, K. M.; Lee, H. J.; Wang, L.; Silverstein, M. C.; Ma'ayan, A. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **2018**, 9 (1), 1366.
- (51) Karlsson, M.; Zhang, C.; Méar, L.; Zhong, W.; Digre, A.; Katona, B.; Sjöstedt, E.; Butler, L.; Odeberg, J.; Dusart, P.; Edfors, F.; Oksvold, P.; von Feilitzen, K.; Zwahlen, M.; Arif, M.; Altay, O.; Li, X.; Ozcan, M.; Mardinoglu, A.; Fagerberg, L.; Mulder, J.; Luo, Y.; Ponten, F.; Uhlen, M.; Lindskog, C. A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **2021**, 7 (31), eabh2169.
- (52) Tabula Sapiens, C.; Jones, R. C.; Karkanas, J.; Krasnow, M. A.; Pisco, A. O.; Quake, S. R.; Salzman, J.; Yosef, N.; Bulthaupt, B.; Brown, P.; Harper, W.; Hemenez, M.; Ponnusamy, R.; Salehi, A.; Sanagavarapu, B. A.; Spallino, E.; Aaron, K. A.; Concepcion, W.; Gardner, J. M.; Kelly, B.; Neidlinger, N.; Wang, Z.; Crasta, S.; Kolluru, S.; Morri, M.; Pisco, A. O.; Tan, S. Y.; Travaglini, K. J.; Xu, C.; Alcántara-Hernández, M.; Almanzar, N.; Antony, J.; Beyersdorf, B.; Burhan, D.; Calcuttawala, K.; Carter, M. M.; Chan, C. K. F.; Chang, C. A.; Chang, S.; Colville, A.; Crasta, S.; Culver, R. N.; Cvijovic, I.; D'Amato, G.; Ezran, C.; Galdos, F. X.; Gillich, A.; Goodyer, W. R.; Hang, Y.; Hayashi, A.; Houshdaran, S.; Huang, X.; Irwin, J. C.; Jong, S.; Juanico, J. V.; Kershner, A. M.; Kim, S.; Kiss, B.; Kolluru, S.; Kong, W.; Kumar, M. E.; Kuo, A. H.; Leykle, R.; Li, B.; Loeb, G. B.; Lu, W.-J.; Mantri, S.; Markovic, M.; McAlpine, P. L.; de Morree, A.; Morri, M.; Mrouj, K.; Mukherjee, S.; Muser, T.; Neuhöfer, P.; Nguyen, T. D.; Perez, K.; Phansalkar, R.; Pisco, A. O.; Puluca, N.; Qi, Z.; Rao, P.; Raquer-McKay, H.; Schaum, N.; Scott, B.; Seddighzadeh, B.; Segal, J.; Sen, S.; Sikandar, S.; Spencer, S. P.; Steffes, L. C.; Subramaniam, V. R.; Swarup, A.; Swift, M.; Travaglini, K. J.; Van Treuren, W.; Trimm, E.; Veizades, S.; Vijayakumar, S.; Vo, K. C.; Vorperian, S. K.; Wang, W.; Weinstein, H. N. W.; Winkler, J.; Wu, T. T. H.; Xie, J.; Yung, A. R.; Zhang, Y.; Detweiler, A. M.; Mekonen, H.; Neff, N. F.; Sit, R. V.; Tan, M.; Yan, J.; Bean, G. R.; Charu, V.; Forgó, E.; Martin, B. A.; Ozawa, M. G.; Silva, O.; Tan, S. Y.; Toland, A.; Vemuri, V. N. P.; Afik, S.; Awaysan, K.; Botvinnik, O. B.; Byrne, A.; Chen, M.; Dehghannasiri, R.; Detweiler, A. M.; Gayoso, A.; Granados, A. A.; Li, Q.; Mahmoudabadi, G.; McGeever, A.; de Morree, A.; Olivieri, J. E.; Park, M.; Pisco, A. O.; Ravikumar, N.; Salzman, J.; Stanley, G.; Swift, M.; Tan, M.; Tan, W.; Tarashansky, A. J.; Vanheusden, R.; Vorperian, S. K.; Wang, P.; Wang, S.; Xing, G.; Xu, C.; Yosef, N.; Alcántara-

- Hernández, M.; Antony, J.; Chan, C. K. F.; Chang, C. A.; Colville, A.; Crasta, S.; Culver, R.; Dethlefsen, L.; Ezran, C.; Gillich, A.; Hang, Y.; Ho, P.-Y.; Irwin, J. C.; Jang, S.; Kershner, A. M.; Kong, W.; Kumar, M. E.; Kuo, A. H.; Leylek, R.; Liu, S.; Loeb, G. B.; Lu, W.-J.; Maltzman, J. S.; Metzger, R. J.; de Morree, A.; Neuhöfer, P.; Perez, K.; Phansalkar, R.; Qi, Z.; Rao, P.; Raquer-McKay, H.; Sasagawa, K.; Scott, B.; Sinha, R.; Song, H.; Spencer, S. P.; Swarup, A.; Swift, M.; Travaglini, K. J.; Trimm, E.; Veizades, S.; Vijayakumar, S.; Wang, B.; Wang, W.; Winkler, J.; Xie, J.; Yung, A. R.; Artandi, S. E.; Beachy, P. A.; Clarke, M. F.; Giudice, L. C.; Huang, F. W.; Huang, K. C.; Idoyaga, J.; Kim, S. K.; Krasnow, M.; Kuo, C. S.; Nguyen, P.; Quake, S. R.; Rando, T. A.; Red-Horse, K.; Reiter, J.; Relman, D. A.; Sonnenburg, J. L.; Wang, B.; Wu, A.; Wu, S. M.; Wyss-Coray, T. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **2022**, 376 (6594), eabl4896.
- (53) Edfors, F.; Danielsson, F.; Hallström, B. M.; Käll, L.; Lundberg, E.; Pontén, F.; Forsström, B.; Uhlén, M. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **2016**, 12 (10), 883.
- (54) Bileschi, M. L.; Belanger, D.; Bryant, D. H.; Sanderson, T.; Carter, B.; Sculley, D.; Bateman, A.; DePristo, M. A.; Colwell, L. J. Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **2022**, 40 (6), 932–937.
- (55) Kustatscher, G.; Grabowski, P.; Schrader, T. A.; Passmore, J. B.; Schrader, M.; Rappsilber, J. Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **2019**, 37 (11), 1361–1371.
- (56) Ochoa, D.; Jarnuczak, A. F.; Viéitez, C.; Gehre, M.; Soucheray, M.; Mateus, A.; Kleefeldt, A. A.; Hill, A.; Garcia-Alonso, L.; Stein, F.; Krogan, N. J.; Savitski, M. M.; Swaney, D. L.; Vizcaino, J. A.; Noh, K.-M.; Beltrao, P. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **2020**, 38 (3), 365–373.
- (57) Bludau, I.; Willems, S.; Zeng, W.-F.; Strauss, M. T.; Hansen, F. M.; Tanzer, M. C.; Karayel, O.; Schulman, B. A.; Mann, M. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.* **2022**, 20 (5), e3001636.
- (58) Keil, B. *Specificity of proteolysis*; Springer Science & Business Media: 1992.
- (59) Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun. Mass Spectrom.* **2000**, 14 (6), 496–502.
- (60) Walmsley, S. J.; Rudnick, P. A.; Liang, Y.; Dong, Q.; Stein, S. E.; Nesvizhskii, A. I. Comprehensive analysis of protein digestion using six trypsin reveals the origin of trypsin as a significant source of variability in proteomics. *J. Proteome Res.* **2013**, 12 (12), 5666–5680.
- (61) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, 7 (1), 300–305.
- (62) Fannes, T.; Vandermarliere, E.; Schietgat, L.; Degroove, S.; Martens, L.; Ramon, J. Predicting tryptic cleavage from proteomics data using decision tree ensembles. *J. Proteome Res.* **2013**, 12 (5), 2253–2259.
- (63) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics* **2005**, 4 (9), 1265–1272.
- (64) Reker, D.; Malmström, L. Bioinformatic challenges in targeted proteomics. *J. Proteome Res.* **2012**, 11 (9), 4393–4402.
- (65) Gao, Z.; Chang, C.; Yang, J.; Zhu, Y.; Fu, Y. AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility. *Anal. Chem.* **2019**, 91 (13), 8705–8711.
- (66) Kailash, V.; Mendoza, L.; Moritz, R. L.; Hoopmann, M. R. SPACEPro: A Software Tool for Analysis of Protein Sample Cleavage for Tandem Mass Spectrometry. *J. Proteome Res.* **2021**, 20 (4), 1911–1917.
- (67) Zhong, X.; Chen, H.; Zare, R. N. Ultrafast enzymatic digestion of proteins by microdroplet mass spectrometry. *Nat. Commun.* **2020**, 11 (1), 1049.
- (68) Wall, M. J.; Crowell, A. M. J.; Simms, G. A.; Liu, F.; Doucette, A. A. Implications of partial tryptic digestion in organic-aqueous solvent systems for bottom-up proteome analysis. *Anal. Chim. Acta* **2011**, 703 (2), 194–203.
- (69) Vaškevičius, M.; Kapočūtė-Dzikiėnė, J.; Šlepikas, L. Prediction of Chromatography Conditions for Purification in Organic Synthesis Using Deep Learning. *Molecules* **2021**, 26 (9), 2474.
- (70) Moruz, L.; Pichler, P.; Stranzl, T.; Mechtler, K.; Käll, L. Optimized nonlinear gradients for reversed-phase liquid chromatography in shotgun proteomics. *Anal. Chem.* **2013**, 85 (16), 7777–7785.
- (71) Trudgian, D. C.; Fischer, R.; Guo, X.; Kessler, B. M.; Mirzaei, H. GOAT—a simple LC-MS/MS gradient optimization tool. *Proteomics* **2014**, 14 (12), 1467–1471.
- (72) Mostovenko, E.; Hassan, C.; Rattke, J.; Deelder, A. M.; van Veelen, P. A.; Palmblad, M. Comparison of peptide and protein fractionation methods in proteomics. *EuPA Open Proteomics* **2013**, 1, 30–37.
- (73) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein pK_a Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput* **2022**, 18 (4), 2673–2686.
- (74) Lindsey, R. K.; Rafferty, J. L.; Eggimann, B. L.; Siepmann, J. I.; Schure, M. R. Molecular simulation studies of reversed-phase liquid chromatography. *Journal of Chromatography A* **2013**, 1287, 60–82.
- (75) Meek, J. L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. U. S. A.* **1980**, 77 (3), 1632–6.
- (76) Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* **2002**, 74 (22), 5826–30.
- (77) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolić, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* **2003**, 75 (5), 1039–48.
- (78) Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W.-J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; Camp, D. G.; Smith, R. D. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* **2006**, 78 (14), 5026–5039.
- (79) Krokhn, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Molecular & Cellular Proteomics* **2004**, 3 (9), 908–919.
- (80) Moruz, L.; Tomazela, D.; Käll, L. Training, selection, and robust calibration of retention time models for targeted proteomics. *J. Proteome Res.* **2010**, 9 (10), S209–S216.
- (81) Ma, C.; Ren, Y.; Yang, J.; Ren, Z.; Yang, H.; Liu, S. Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **2018**, 90 (18), 10881–10888.
- (82) Zeng, W.-F.; Zhou, X.-X.; Willems, S.; Ammar, C.; Wahle, M.; Bludau, I.; Voytik, E.; Strauss, M. T.; Mann, M. AlphaPeptDeep: A modular deep learning framework to predict peptide properties for proteomics. *Nat. Commun.* **2022**, 13, 7238.
- (83) Maboudi Afkham, H.; Qiu, X.; The, M.; Käll, L. Uncertainty estimation of predictions of peptides' chromatographic retention times in shotgun proteomics. *Bioinformatics* **2016**, 33 (4), 508–513.
- (84) Van Puyvelde, B.; Willems, S.; Gabriels, R.; Daled, S.; De Clerck, L.; Vande Castele, S.; Staes, A.; Impens, F.; Deforce, D.; Martens, L.; Degroove, S.; Dhaenens, M. Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *Proteomics* **2020**, 20 (3–4), e1900306.
- (85) Searle, B. C.; Swearingen, K. E.; Barnes, C. A.; Schmidt, T.; Gessulat, S.; Küster, B.; Wilhelm, M. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* **2020**, 11 (1), 1548.

- (86) Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R. Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (10), 1835–1843.
- (87) Argentini, A.; Goeminne, L. J. E.; Verheggen, K.; Hulstaert, N.; Staes, A.; Clement, L.; Martens, L. moFF: a robust and automated approach to extract peptide ion intensities. *Nat. Methods* **2016**, *13* (12), 964–966.
- (88) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **2012**, *12* (8), 1111–1121.
- (89) Mohammed, Y.; Palmblad, M. Visualization and application of amino acid retention coefficients obtained from modeling of peptide retention. *J. Sep. Sci.* **2018**, *41* (18), 3644–3653.
- (90) Dodds, J. N.; Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (11), 2185–2195.
- (91) Dodds, J. N.; May, J. C.; McLean, J. A. Correlating Resolving Power, Resolution, and Collision Cross Section: Unifying Cross-Platform Assessment of Separation Efficiency in Ion Mobility Spectrometry. *Anal. Chem.* **2017**, *89* (22), 12176–12184.
- (92) Broeckling, C. D.; Yao, L.; Isaac, G.; Gioioso, M.; Ianchis, V.; Vissers, J. P. C. Application of Predicted Collisional Cross Section to Metabolome Databases to Probabilistically Describe the Current and Future Ion Mobility Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (3), 661–669.
- (93) Zhou, Z.; Xiong, X.; Zhu, Z.-J. MetCCS predictor: a web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics* **2017**, *33* (14), 2235–2237.
- (94) Chang, C.-H.; Yeung, D.; Spicer, V.; Ogata, K.; Krokhin, O.; Ishihama, Y. Sequence-Specific Model for Predicting Peptide Collision Cross Section Values in Proteomic Ion Mobility Spectrometry. *J. Proteome Res.* **2021**, *20*, 3600.
- (95) Steigenberger, B.; van den Toorn, H. W. P.; Bijl, E.; Greisch, J.-F.; Räther, O.; Lubeck, M.; Pieters, R. J.; Heck, A. J. R.; Scheltema, R. A. Benefits of Collisional Cross Section Assisted Precursor Selection (caps-PASEF) for Cross-linking Mass Spectrometry. *Mol. Cell Proteomics* **2020**, *19* (10), 1677–1687.
- (96) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. P. Collision cross sections for structural proteomics. *Structure* **2015**, *23* (4), 791–799.
- (97) Turzo, S. M. B. A.; Seffernick, J. T.; Rolland, A. D.; Donor, M. T.; Heinze, S.; Prell, J. S.; Wysocki, V. H.; Lindert, S. Protein shape sampled by ion mobility mass spectrometry consistently improves protein structure prediction. *Nat. Commun.* **2022**, *13* (1), 4377.
- (98) O'Leary, M. H. Carbon isotope fractionation in plants. *Phytochemistry* **1981**, *20* (4), 553–567.
- (99) Lederberg, J. DENDRAL-64: a system for computer construction, enumeration and notation of organic molecules as tree structures and cyclic graphs, Part I: notational algorithm for tree structures. *NASA Scientific and Technical Aerospace Report*; N65-13158; NASA: 1964.
- (100) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.
- (101) Tiwary, S.; Levy, R.; Gutenbrunner, P.; Salinas Soto, F.; Palaniappan, K. K.; Deming, L.; Berndt, M.; Brant, A.; Cimermanic, P.; Cox, J. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **2019**, *16* (6), 519–525.
- (102) Zhou, X. X.; Zeng, W. F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S. M.; Zhang, Z. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **2017**, *89* (23), 12690–12697.
- (103) Zimmer, D.; Schneider, K.; Sommer, F.; Schroda, M.; Mühlhaus, T. Artificial Intelligence Understands Peptide Observability and Assists With Absolute Protein Quantification. *Front. Plant Sci.* **2018**, *9*, 1559.
- (104) Pino, L. K.; Searle, B. C.; Yang, H.-Y.; Hoofnagle, A. N.; Noble, W. S.; MacCoss, M. J. Matrix-Matched Calibration Curves for Assessing Analytical Figures of Merit in Quantitative Proteomics. *J. Proteome Res.* **2020**, *19* (3), 1147–1153.
- (105) Nielsen, M. L.; Savitski, M. M.; Kjeldsen, F.; Zubarev, R. A. Physicochemical properties determining the detection probability of tryptic peptides in Fourier transform mass spectrometry. A correlation study. *Anal. Chem.* **2004**, *76* (19), 5872–5877.
- (106) 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* **2007**, *8* (Suppl 7), S23.
- (107) Tang, H.; Arnold, R. J.; Alves, P.; Xun, Z.; Clemmer, D. E.; Novotny, M. V.; Reilly, J. P.; Radivojac, P. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **2006**, *22* (14), e481–8.
- (108) Dincer, A. B.; Lu, Y.; Schweppe, D. K.; Oh, S.; Noble, W. S. Reducing Peptide Sequence Bias in Quantitative Mass Spectrometry Data with Machine Learning. *J. Proteome Res.* **2022**, *21* (7), 1771–1782.
- (109) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature* **2011**, *473* (7347), 337–342.
- (110) Dai, C.; Füllgrabe, A.; Pfeuffer, J.; Solovyeva, E. M.; Deng, J.; Moreno, P.; Kamatchinathan, S.; Kundu, D. J.; George, N.; Fexova, S.; Grüning, B.; Föll, M. C.; Griss, J.; Vaudel, M.; Audain, E.; Locard-Paulet, M.; Turewicz, M.; Eisenacher, M.; Uszkoreit, J.; Van Den Bossche, T.; Schwämmle, V.; Webel, H.; Schulze, S.; Bouyssié, D.; Jayaram, S.; Duggineni, V. K.; Samaras, P.; Wilhelm, M.; Choi, M.; Wang, M.; Kohlbacher, O.; Brazma, A.; Papatheodorou, I.; Bandeira, N.; Deutsch, E. W.; Vizcaino, J. A.; Bai, M.; Sachsenberg, T.; Levitsky, L. I.; Perez-Riverol, Y. A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat. Commun.* **2021**, *12* (1), 5854.
- (111) Shrestha, D. L.; Solomatine, D. P. Machine learning approaches for estimation of prediction interval for the model output. *Neural Netw.* **2006**, *19* (2), 225–235.
- (112) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15* (2), 906–915.
- (113) Moon, J.; Kim, J.; Shin, Y.; Hwang, S. Confidence-aware learning for deep neural networks. In *International conference on machine learning*; PMLR: 2020; pp 7034–7044.
- (114) Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarek, V.; Nahavandi, S. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* **2021**, *76*, 243–297.
- (115) The, M.; Käll, L. Integrated Identification and Quantification Error Probabilities for Shotgun Proteomics. *Mol. Cell Proteomics* **2019**, *18* (3), 561–570.
- (116) McKennan, C.; Sang, Z.; Shi, Y. A novel framework to quantify uncertainty in peptide-tandem mass spectrum matches with application to nanobody peptide identification. *arXiv* **2021**, DOI: 10.48550/arXiv.2110.07818.
- (117) Listgarten, J.; Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell Proteomics* **2005**, *4* (4), 419–434.
- (118) Jaitly, N.; Monroe, M. E.; Petyuk, V. A.; Clauss, T. R. W.; Adkins, J. N.; Smith, R. D. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.* **2006**, *78* (21), 7397–7409.

- (119) The, M.; Käll, L. Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. *Nat. Commun.* **2020**, *11* (1), 3234.
- (120) Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of Machine Learning Research* **2016**, *48*, 1050–1059.
- (121) Bielow, C.; Aiche, S.; Andreotti, S.; Reinert, K. MSSimulator: Simulation of mass spectrometry data. *J. Proteome Res.* **2011**, *10* (7), 2922–2929.
- (122) Kösters, M.; Leufken, J.; Leidel, S. A. SMITER-A Python Library for the Simulation of LC-MS/MS Experiments. *Genes (Basel)* **2021**, *12* (3), 396.
- (123) Leeming, M. G.; Ang, C.-S.; Nie, S.; Varshney, S.; Williamson, N. A. Simulation of mass spectrometry-based proteomics data with Synthedia. *Bioinformatics Advances* **2023**, *3* (1). DOI: 10.1093/bioadv/vbac096.
- (124) Awan, M. G.; Saeed, F. MaSS-Simulator: A Highly Configurable Simulator for Generating MS/MS Datasets for Benchmarking of Proteomics Algorithms. *Proteomics* **2018**, *18* (20), e1800206.
- (125) Bik, E. M.; Casadevall, A.; Fang, F. C. The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. *mBio* **2016**, *7* (3), e00809–16.
- (126) Bradshaw, M. S.; Payne, S. H. Detecting fabrication in large-scale molecular omics data. *PLoS One* **2021**, *16* (11), e0260395.
- (127) Erratum: Liesz.; et al. DAMP Signaling Is a Key Pathway Inducing Immune Modulation after Brain Injury. *J. Neurosci.* **2019**, *39* (27), 5419.
- (128) Further update on research misconduct investigation, University of Liverpool News, 17 August, 2020.
- (129) Liesz, A.; Dalpke, A.; Mracsko, E.; Roth, S.; Zhou, W.; Yang, H.; Na, S.-Y.; Akhisaroglu, M.; Fleming, T.; Eigenbrod, T.; Nawroth, P. P.; Tracey, K. J.; Veltkamp, R. DAMP signaling is a key pathway inducing immune modulation after brain injury. *J. Neurosci.* **2015**, *35* (2), 583–598.
- (130) Diekmann, A. Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics* **2007**, *34* (3), 321–329.
- (131) Zipf, G. K. *The psycho-biology of language: An introduction to dynamic philology*; Routledge: 2013.
- (132) Deutsch, E. W.; Bandeira, N.; Sharma, V.; Perez-Riverol, Y.; Carver, J. J.; Kundu, D. J.; García-Seisdedos, D.; Jarnuczak, A. F.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; Hermjakob, H.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaino, J. A. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* **2019**, *48* (D1), D1145–D1152.
- (133) Fisher, R. *The use of multiple measurements in taxonomic problems*, *Annual Eugenics*, *7*, Part II, 179–188 (1936); also in *Contributions to Mathematical Statistics*; John Wiley: NY; 1950.
- (134) Dawson, R. J. M., The "unusual episode" data revisited. *Journal of Statistics Education* **1995**, *3* (3).
- (135) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; Yu, P.; Schlegel, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **2017**, *14* (3), 259–262.
- (136) Rehfeldt, T. G.; Krawczyk, K.; Bøgebjerg, M.; Schwämmle, V.; Röttger, R. MS2AI: Automated repurposing of public peptide LC-MS data for machine learning applications. *Bioinformatics* **2022**, *38* 875.877
- (137) Rehfeldt, T. G.; Gabriels, R.; Bouwmeester, R.; Gessulat, S.; Neely, B. A.; Palmblad, M.; Perez-Riverol, Y.; Schmidt, T.; Vizcaino, J. A.; Deutsch, E. W. ProteomicsML: An Online Platform for Community-Curated Datasets and Tutorials for Machine Learning in Proteomics. *J. Proteome Res.* **2023**, *22*, 632–636.
- (138) Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, 2017.
- (139) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9* (2), 513–530.
- (140) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaino, J. A. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research* **2022**, *50* (D1), D543–D552.
- (141) Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst* **2018**, *7* (4), 412–421.e5.
- (142) Foster, J. M.; Degroove, S.; Gatto, L.; Visser, M.; Wang, R.; Griss, J.; Apweiler, R.; Martens, L. A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* **2011**, *11* (11), 2182–2194.
- (143) Maslow, A. H. *The Psychology of Science: A Reconnaissance*; Harper & Row: 1966.