Editorial summary: Proteomics is being transformed by deep-learning methods that predict peptide fragmentation spectra.

Prediction of peptide mass spectral libraries with machine learning

Jürgen Cox[1,2,*]

[1]Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.
[2]Department of Biological and Medical Psychology, University of Bergen, Jonas Liesvei 91, 5009 Bergen, Norway.
[*]cox@biochem.mpg.de

Abstract

The recent development of machine-learning methods to identify peptides in complex mass-spectrometric data constitutes a major breakthrough in proteomics. Longstanding methods for peptide identification, such as search engines and experimental spectral libraries, are being superseded by deep-learning models that allow the fragmentation spectra of peptides to be predicted from their amino-acid sequence. These new approaches, including recurrent neural networks and convolutional neural networks, use predicted *in silico* spectral libraries rather than experimental libraries to achieve higher sensitivity and/or specificity in the analysis of proteomics data**.** Notably, machine learning is galvanizing applications that involve large search spaces, such as immunopeptidomics and proteogenomics. Current challenges in the field include the prediction of spectra for peptides with posttranslational modifications and for cross-linked pairs of peptides. Permeation of machine learning-based spectral prediction into search engines and spectrum-centric data-independent acquisition workflows for diverse peptide classes and measurement conditions will continue to push sensitivity and dynamic range in proteomics applications in the coming years.

Introduction

Shotgun proteomics[1–4] is a technique to identify and quantify proteins in samples of interest (Fig, 1). The approach includes two main steps. First, proteins are digested to peptides by proteases, and second, the peptides are fragmented in the mass spectrometer, which results in fragmentation spectra. Because of the oligomeric structure of peptides and the dominance of bond breakage in their backbone, the fragmentation spectra display regularities[5] that can be exploited to determine their sequence of amino acids[6] and covalent modifications of the amino acids. Thus, the interpretation of peptide spectra is somewhat different compared with molecules that lack a repetitive structure, such as metabolites[7]. With knowledge of the physical method of fragmentation — such as collision induced dissociation[8,9], higher-energy collisional dissociation[10] or electron transfer dissociation[11] — the masses of the dominant peptide fragments can be easily calculated from the sequence. However, it is non-trivial to predict the relative intensities of fragment peaks in the spectrum, or in some cases their absence from the spectrum, which are both determined by quantum chemistry[12]. The peptide search engines[13–15] that are traditionally used in shotgun proteomics to identify peptides generally ignore intensity information beyond simplified rules[16,17]. Although these tools have been successfully applied for many years, the intensity pattern carries information that can be used to improve the sensitivity and specificity of the peptide identification process[18–20].

One method for making use of the intensity information is to assemble libraries directly from previously measured spectra[21–24] and apply them to the analysis of the sample of interest. This approach has the advantage that it is hypothesis-free regarding the content of the spectra. In principle, it can accommodate non-standard peaks that do not belong to any of the standard fragment ion series[25] (Fig. 1), which are not considered in most prediction approaches available today. The disadvantage is that any new peptides in the sample, for which no library spectrum has been acquired, that would be lost in the analysis. An alternative approach, which, however, is prone to lose new peptides in the analysis as well, is to acquire project-specific libraries. Generating such libraries adds substantial measurement effort to the project and is usually restricted to applications in which the benefit of increased sensitivity strongly outweighs the shortcoming of the peptide space being limited by the library content.

The limitations of these approaches for leveraging intensity information — both the failure to detect novel peptides and the additional measurement effort of generating project-specific libraries — would be overcome if peak intensities in fragmentation spectra could be predicted quickly and precisely from amino acid sequences. The first attempts at doing this date back nearly 20 years, using either a decision tree[26] or a single hidden layer neural network[27]. In a recent breakthrough, deep learning methods[28,29] have begun to predict peptide fragmentation spectra from amino acid sequence with near experimental accuracy[18,30,31]. This review focuses on machine-learning methods for accurate prediction of spectral libraries . Other recent reviews have covered more generally how deep learning is applied to proteomics[32–34].

The first section introduces key concepts in machine learning and deep learning. The second section discusses the requirements for training data used in spectral prediction and methods that predict either selected ion series or the full spectrum. The third section covers prediction of spectra for cross-linked peptides and peptides with post-translational modifications; in these applications the size of the available training data is currently much smaller, posing additional challenges. Predicted spectral libraries are beneficial in both data-dependent acquisition[35–41] (DDA) and data-independent acquisition[42] (DIA) methods. The fourth section explains how intensity-based rescoring in DDA is facilitating applications in immunoproteomics and proteogenomics, two areas with a substantially larger search space compared with standard proteomics. The fifth section discusses applications in DIA. Whereas DDA isolates a single molecular species for fragmentation, DIA fragments many peptides simultaneously (Fig. 2), generating spectra that are much more complex. The added information from the library spectrum intensities is particularly beneficial to deconvolute contributions from different peptides. This section compares DIA experiments analyzed with experimental versus predicted libraries and ends with a short overview of recent specialized spectrum prediction tools for DIA applications. Finally, a concluding section discusses promising avenues for further improving machine-learning tools for spectral prediction.

Machine learning and deep learning approaches

Fragmentation spectrum prediction is a supervised learning problem in which the spectrum is predicted from the peptide sequence and models are trained on sets of peptide sequences plus metadata, e.g. peptide charge or collision energy, as the input variables, and the fragment intensities, for instance of the *y*- and *b*-series ions, as the output variables (Fig. 3a). There is a

large variety of regression methods available, including tree-based models[43,44] such as random forest regression[45] and XGBoost[46], support vector regression[47,48] and neural networks. Neural networks are frequently used in spectrum prediction due to their superior performance.

Neural networks are machine-learning methods that are roughly modeled on how neurons integrate signals in the brain. Directed connections carrying weights indicate which other neurons a given neuron can communicate with. A single idealized computational neuron (Fig. 3b) has incoming and outgoing connections, where incoming connections can have an excitatory or an inhibitory effect, depending on whether their weights are positive or negative. The output is calculated as a nonlinear ('activation') function applied to the sum of the input signals, and sent to the next neurons via the outgoing connections. Neurons are typically arranged in layers with the multilayer perceptron (Fig. 3c) as the prototypic example of a feedforward network. The weights are determined during training, in which examples are presented for which the outcome (the spectrum) is known. A loss function measures the discrepancy between the true outcome and the current prediction. This prediction error is minimized in a strategy called back-propagation[49]. Deep neural networks have architectures, i.e. content and connectivity of computational neurons, that are sufficiently complex to represent data in a hierarchy of concepts, where complex representations are composed of simpler ones. In the example of a multilayer perceptron in Fig. 3c, the presence of multiple hidden layers, i.e. layers of neurons, that are not directly connected to the input or the output, would allow for such a hierarchical representation. Learning the representation of higher-level concepts in the original, primitive data is one of the hallmarks of deep learning, in contrast to simpler, conventional machine learning algorithms that are applied to hand-crafted features of pre-processed data. This approach is very practical and powerful, since no extensive domain knowledge is required; however, it can come at the expense of an increased need in the number of training instances and computation time, compared with conventional machine learning on extracted features.

One specific class of neural networks, recurrent neural networks[49,50] (RNN), has turned out to be very useful for fragmentation spectrum prediction. They are designed to process sequential data and can be applied to sequences of variable length, which makes them particularly applicable to peptides. Bidirectional RNNs[51] combine two RNNs, one for each direction

along the sequence, to take into account that the frequency of a certain bond breakage depends on the sequence context before and after that bond. For some peptide bonds, their propensity to break is determined mostly by the local molecular environment, whereas for others more distant sequence properties are relevant. Gated RNNs have been developed to deal with multiple causal distance scales in the sequence. The two main types of gated RNNs, long short-term memory[52,53] (LSTM) and gated recurrent units[54] (GRUs) have both been applied to fragmentation spectrum prediction. Also, convolutional neural networks[55] (CNNs), which have traditionally been used for tasks in image classification and recognition have been applied to spectra.

Transfer learning[56] (Fig. 3d) is a technique wherein parts of a trained model are re-used in a model with a different but related task which is then fine-tuned by a smaller number of training instances than would have been needed if the model had been trained from scratch. This technique can be useful for applications in which spectra are predicted for specialized technological or biological contexts, for which one can borrow parts of trained models from a more generic context. For instance, a model trained on a large dataset of unmodified peptides can be partially transferred to a model of peptides carrying posttranslational modifications which is subsequently trained on a smaller dataset. Finally it is of interest that computational methods such as shapely additive explanations[57] (SHAP) and integrated gradients[58] are available for the attribution of input feature ranges to the prediction outcomes for a particular instance. In image recognition, for instance, these methods can indicate pixel ranges in an image that are most responsible for a certain decision. Similarly, in spectra they can provide information on the sequence regions that are most contributing to the determination of a fragment ion intensity[18].

Spectral data

Fragmentation spectra can be predicted in two ways, either by focusing on pre-defined ion series types, e.g. *y*- and *b*-type ions, whose masses are directly calculable from the input sequence and whose intensities are to be predicted, or by predicting the full spectrum without referring to ion series annotation. Crucial for training a predictive model is a dataset of examples for which the input and the output is known. Such a ground truth dataset can be obtained from synthetic peptides with defined sequences[59,60] which then undergo mass spectrometric analysis. This approach has the advantage that the entire composition of the

peptide mixture is known. However, spectra obtained from such measurements do not reflect the composition of a real sample, since they cover only a limited set of peptides, and substantial efforts are required for synthesis and analysis. More often one makes use of existing DDA datasets, deposited in public raw data repositories[61–67]. In this case it is ensured that peptides are correctly identified up to a selectable false discovery rate[68,69] (FDR) and can therefore serve as a quasi-ground truth. Optionally, further thresholds on additional quality parameters such as the search engine score can be applied. For approaches predicting the full spectrum, re-analyzing complex proteomics data to train the model has the complication that peaks can have resulted from co-fragmented peptides, which either would have to be reduced by spectral clustering[70,71], or by a threshold on a measure for co-fragmentation[72,73]. Otherwise, the machine learning model will have the additional task of identifying features that are present due to co-fragmentation. In contrast, in approaches that predict only the intensities of ion series, the contamination effect of co-fragmented peptides is expected to be minor even in complex proteome samples.

To determine the performance of a machine learning model, the available data needs to be split into a training, a validation and a test data set. The training and validation sets are used for model building, whereas the test set is entirely excluded from this process but is afterwards used to assess the performance of the model in terms of predictive accuracy in an unbiased way (Fig. 3e). For model building, the training dataset is used to determine the parameters of the model, i.e. the weights and biases of a neural network, and the validation set is used to tune the model's hyperparameters and to avoid overfitting in this process. To judge the accuracy of a prediction, one needs a spectral similarity measure[74–77] which quantifies how close the predicted spectrum is to the experimental one, examples include the Pearson correlation between the spectral intensities or the spectral contrast angle[78]. Calculating the similarity measure for all predictions on the test set elements results in a histogram, which can be used to calculate the average accuracy, confidence intervals or a box plot for the whole population of predictions. In cases where the available data are limited, one can use cross validation (Fig. 3f) to increase the statistics of the histogram of accuracies.

The predictive performance of a model depends on the number of available training instances. If the training set is too small, the full potential of the approach might not have been reached and one would need to obtain more instances to reach the plateau of asymptotic

performance (Fig. 3g). In practice, it is important how the model performs with a limited number of training instances, since the number of available spectra in a given technological or biological setting might be restricted. There is a practical limit set to the prediction accuracy given by how similar technical replicates of MS/MS spectra are for the same peptide and same values of metadata parameters[79].

Ion series intensity prediction

Most of the popular deep learning models for ion series intensity prediction use RNNs, which have been realized in pDeep[30,80], DeepMass:Prism[18], Prosit[31] and by Guan et al[81], but CNNs were also used[82]. Prosit is based on a GRU, whereas the other RNN based models use LSTM layers. As an example, the architecture of DeepMass:Prism (Fig 4a) is described in more detail. It uses the encoder-decoder architecture[83] which has been developed in the context of machine translation, e.g. for turning German sentences into English. The encoder part takes a variable length peptide sequence as input and transforms it into a fixed length representation, which is achieved by three LSTM layers. Together with values of metadata parameters, such as charge or fragmentation type, a decoder consisting of a multilayer perceptron generates the 'translated' sequence of ion series intensities. Outputs include y and b ions as well as peaks resulting from losses of H_2O and NH_3. Prosit also follows an encoder-decoder architecture but has slight differences in its construction, as it takes the normalized collision energy as an additional metadata parameter input.

Conventional machine learning has been applied to ion series intensity prediction as well. These methods can be sub-divided into fixed length and window-based approaches. In the former, which is implemented in MS2PIP[84–86], a separate model is trained for every possible peptide length (Fig 4b). Thus, there is no synergy from peptides of different lengths as it is the case for RNNs. Since there is no complication from variable-length inputs, in principle any conventional machine learning algorithm could be used with random forests as first choice[84]. For window-based methods such as wiNNer[18], which is categorized as deep learning since its neural network contains multiple hidden layers, peptides of different lengths contribute to the same model (Fig 4c). The model predicts the peak heights relative to the highest peak in the spectrum for the ions formed by the breakage of one peptide bond at a time. The feature space is of fixed length and can be thought of as representing a sequence window around the currently considered bond plus some additional features. Features include

one-hot encoded amino acids in the sequence window centered on the peptide bond under consideration, length of the peptide, distances (number of residues) to the C- and N-terminus, one-hot encoded amino acids at the termini plus the values of metadata parameters that were also fed into the RNN-based models. Multiple instances of window-based training data will be created from one peptide by sliding the window along the sequence. Several other approaches also belong in this category since their prediction focuses on one peptide bond at a time and the features are recruited partially from the amino acids around that bond[26,27,87–89] with a small window size. Although the prediction accuracy of window-based prediction is usually lower compared with RNN based prediction, it can come close and it has the potential merits of lesser need in the number of training instances and decreased computational complexity[18,90].

Full spectrum prediction

A CNN-based architecture was developed for the prediction of full spectra including also non-backbone ions[91]. The method does not rely on peak annotation, instead it uses a binned m/z range up to 2000 Da with a bin width of 0.1 resulting in a vector of 20,000 dimensions as a target for predicted intensities. A one-hot encoding for the input sequence is used to predict doubly and triply charged unmodified higher energy collisional dissociation (HCD) spectra, for which many training instances are available. About 1.5 million spectra were needed to reach saturation in prediction accuracy. Since much fewer charge one and four HCD spectra were available for training, multitask learning[92], in which multiple learning tasks are addressed simultaneously to benefit from commonalities, was applied in the prediction of less frequent charge states. An auxiliary prediction task, which is the precursor charge prediction, is integrated into the model as a focusing method to avoid catastrophic forgetting[93]. The prediction of electron transfer dissociation (ETD) spectra was enabled by similar integration with the HCD model by including a pseudo-predictor for the fragmentation type. Future work in full spectrum prediction could include the extension to other fragmentation methods as, for instance, electron transfer/high-energy collision dissociation[94] (EThcD) or ultraviolet photodissociation[95] (UVPD) which are less well understood. Furthermore, the applications of feature attribution methods might shed light onto the mechanisms behind the generation of non-backbone ions. For the assembly of the training data, care needs to be taken to prevent excess of fragments originating from co-fragmented peptides.

Modified and cross-linked peptides

Posttranslational modifications (PTMs) are covalent modifications to proteins that can occur on the amino acid side chains or on the termini. Their presence changes the masses of the ion series members and can also have profound influences on the peak intensities. Furthermore, they can give rise to additional fragments, due to modification-specific neutral losses. For instance, a phosphorylated serine or threonine can produce an additional ion series caused by the loss of H_3PO_4 whereas an immonium ion peak indicates the presence of a phosphorylated tyrosine. Transfer learning was used to modify pDeep2[96] to predict spectra containing modifications. The model was first pretrained on a large dataset of spectra from unmodified peptides. The full model consists of an input layer, two bi-directional LSTM layers and an output layer which was augmented with nodes representing *b*- and *y*-ions caused by the PTM neutral loss. In the transfer learning step, only the first LSTM layer and the output layer are fine-tuned, while the rest of the model is frozen in its pre-trained state. It was found that in particular when only a small number of spectra carrying the PTM are available, the performance of the transfer learned mode is better than the performance of a model trained from scratch[96]. For phosphorylation analysis, the investigators found their prediction of fragment ion intensities of the H_3PO_4 loss to be helpful for site localization. The input sequence features are represented by a 20-dimensional one-hot encoding vector per amino acid plus another vector per amino acid representing the modification in case there is one present. This latter vector is filled with counts of atom types occurring in the modification, thereby encoding its atomic composition. A similar model has recently been used for retention time prediction of modified peptides[97]. The representation can likely be improved in the future, since it cannot represent complex PTMs such as glycosylation[98] adequately, does not distinguish isomers and inherently interpolates between atomic compositions of modifications, which is likely not optimal for representing chemical properties. Fragment spectrum prediction is particularly important for the site localization of PTMs such as phosphorylation when working in a spectral library context[99]. DeepPhospho[100] is another deep learning model that integrates spectral library prediction into a DIA workflow by using a transformer network for the prediction of peptide fragmentation patterns.

Another class of peptides for which specialized methodology is needed for spectrum prediction is produced in cross-linking (XL) mass spectrometry[101]. Here, pairs of peptides are

produced that are covalently connected by a linker that joins two amino acids, one from each of the two peptides. The fragmentation patterns of each of the peptides are influenced by the presence of the other peptide, which makes their prediction harder than for linear peptides. Some of the fragments include the linker and the respective other peptide, which makes them heavier and higher charged on average. Less data are available for cross-linked peptides compared with linear peptides and they show high diversity due to many available cross-linking reagents. Cross linkers can be either cleavable by mass spectrometry or non-cleavable resulting in two different types of fragmentation spectra. pDeepXL[102] is a deep neural network that was trained separately on cleavable and non-cleavable XL data resulting in two prediction models, which are based on transfer learning. Future iterations of deep learning architectures possibly together with retention time predictors for cross-linked peptides[103] are likely to improve the sensitivity of XL search engines[104] when being integrated into their scores.

DDA applications

An important application of accurate fragment spectrum intensity prediction is its use for improving the matching of experimental spectra to peptide candidates. In DDA, the peptide database search engine decides for each given fragmentation spectrum, which among usually several candidates constitutes the best peptide spectrum match (PSM). An overall improvement in the correctness of PSM assignments results in better sensitivity, specificity, or both. Early attempts at intensity integration[105–107] have demonstrated that this is feasible in principle. Recently it was shown[18,108] that by using intensity information, an additional increase in correctness of assignments can be achieved in standard proteome searches against a species-specific sequence database from *homo sapiens* UniProt[109] protein sequences, which contains all tryptic peptides up to a few missed cleavages. One approach directly integrated the intensity information into the Andromeda search engine score[18] (Fig. 5a) and the other used percolator[110,111] for the integration of spectral comparison features with the MS-GF+[112] search engine score[108]. The improvement in sensitivity is q-value (or PSM FDR) dependent and is higher at small q-values. At the standard FDR of 1%, the improvement with deep learning predictions was around 4%. Although the increase in identifications for standard proteomes is only moderate, it is expected that in larger search spaces the benefit of intensity prediction is higher, since on average more potential PSMs exist per precursor mass within a certain tolerance window, among which the correct one needs to be found. Applications with

larger peptide search spaces include immunopeptidomics[113], proteogenomics[114] and metaproteomics[115].

Immunopeptidomics focusses on peptides bound to human leukocyte antigens (HLA), which are generated by proteasomal degradation of intracellular proteins, followed by relocation to the cell surface[116,117]. Defining the HLA peptidomes presented on cancer cells is an intensely studied area of biomedical research, as these peptides provide targets for therapeutic intervention[118,119]. In contrast to proteins, which have to be digested by a specific protease for shot-gun proteomics, HLA peptides can be directly measured by mass spectrometry[120,121], which comes with the challenge of an increased search space due to unspecific cleavage. Furthermore, the rules governing fragmentation differ from those for tryptic peptides; therefore, models for prediction of HLA peptide fragmentation need to be trained extensively also on non-tryptic peptides. Deep learning based intensity prediction was used to improve peptide identification in immunopeptidomics[20,122]. A new Prosit model was trained with more than 300,000 synthetized peptides representing HLA class I and II ligands and cleavage products of the proteases AspN and LysN, allowing the accurate prediction of fragment ion spectra for tryptic and non-tryptic peptides[20]. The researchers reprocessed a dataset consisting of HLA class I peptides from 95 monoallelic cell lines[123] with MaxQuant[124,125] and PSMs were re-scored by integrating fragment intensity predictions. After reprocessing, a 1.5-fold improvement across cell lines in terms of identified peptides was achieved on average (Fig. 5b). Re-scoring with the integrated intensity prediction was also applied to investigate the extent of proteasomal splicing[126,127]. These results suggest that 87% of the proposed proteasomal spliced HLA peptides[126–128] might be incorrect. Many of these did not remain confident after predicted intensity-based rescoring since an equally good or better match with a canonical (non-spliced) peptide was found. In conclusion, the integration of intensity prediction is clearly beneficial for the analysis of HLA ligands.

Proteogenomics[114] is the study of the proteome with the aid of genomic or transcriptomic sequences that allow for the identification of peptides that are not part of the reference proteome sequences. The *in-silico* translation of this extended sequence space leads to an inflation of the peptide search space that has to be taken into consideration when identifying the best PSM for a spectrum. The extent of the search space inflation depends on the scientific question and can range from the inclusion of untranslated regions of transcripts (3’

or 5' UTRs) to the six-frame translation of the whole genome. Proteogenomics also benefits from an integration of predicted spectral intensities through a rescoring of PSMs in a percolator-based approach[19]. Proteogenomics search spaces were generated with ribosomal profiling[129] and with a three-frame translation database based on RNA-seq using nanopores[130]. The latter resulted in an over 50-fold sequence database size growth with an associated 20-fold amino acid content increase. An improvement in the number of identifications was achieved over the whole range of PSM q-values (Fig. 5c), with an increase in identifications of around 6% at the default q-value cutoff of 0.01.

Although the methods that were applied to standard proteomes, proteogenomics and immunopeptidomes differ and are not directly comparable, the results indicate that the improvement is by far the largest in immunopeptidomics, suggesting that the presence of non-tryptic peptides is a more important factor than the size of the search space. Another promising application of deep learning to the peptide identification problem is DeepMatch[131], which circumvents the prediction of spectra and directly predicts PSM scores. Although the approach showed promising results in terms of identification rates, its computational demands turned out to be too high for it to be integrated into regular peptide search engines.

DIA applications

DIA data analysis workflows can be subdivided into spectrum-centric and peptide-centric approaches. Spectrum-centric software tools[132–134] assemble pseudo-DDA spectra from the precursor and fragment features of the DIA data, which are then submitted to conventional search engines. In the peptide-centric approach dedicated spectral libraries are used to query the DIA samples for the peptides represented by the library spectra. Thus, the peptide-centric approach can directly benefit from library prediction. Several peptide-centric software frameworks have been developed[36,135–143] and in principle all of them can be operated with predicted libraries. For standard proteomics samples of a single species without additional enrichments, e.g. for phosphorylation, the use of unbiased full proteome *in silico* predicted libraries for trypsin digestion were found to be feasible and beneficial[143]. Furthermore, error rates on protein identifications are under good statistical control, even when using such large *in silico* libraries[143]. Fig. 6a shows principal component analysis (PCA) results of tissue samples of different cancers measured with DIA using cancer type-specific measured libraries. In Fig. 6b the same data is analyzed in MaxDIA with an in silico predicted library

containing all human tryptic peptides including up to one missed cleavage, which resulted in more protein groups and better separation of tumor types in the PCA. In another example, HEK cell lysate was high-pH reversed-phase peptide fractionated and DIA samples were measured. These were analyzed using three libraries, single-shot and fractionated measured DDA libraries and an *in-silico* full proteome library (Fig. 6c). Also here the predicted library outperforms the measured libraries.

It is of interest how the DIA performance depends on the predictor that is used for generating the *in-silico* library. The same DIA analysis was performed with full proteome spectral libraries predicted by Prosit, DeepMass:Prism and wiNNer (Fig. 6d). The number of proteins and of peptides identified is very similar between the three library generation methods, with a large overlap, implying that the simpler, and hence faster, wiNNer model can be used in place of the RNN-based models without substantial drawbacks.

Several specialized spectrum prediction tools for DIA applications have recently been developed: Predicted peptide libraries can be refined with empirical data [144], or hybrid spectral libraries can be generated that supplement an experiment-derived library with a protein family-targeted in-silico library[145]. DeepDIA[146] uses instrument-specific models and peptide detectability prediction. MSLibrarian[147] optimizes predicted spectral libraries by the integrated usage of spectrum-centric DIA data interpretation[132] to inform and calibrate the *in silico* predicted library and analysis approach.

Conclusion

The predictive accuracy of current spectral library prediction tools is advancing DDA and DIA data analysis. Rescoring of PSMs in DDA is improving their sensitivity-specificity characteristics, in particular for non-tryptic peptides. DIA data analysis can now be routinely performed based on unbiased full proteome prediction of spectral libraries, eliminating the need for measuring project-specific libraries. Despite this progress, proteomics still faces challenges regarding sensitivity. Although cellular proteomes can be routinely quantified with adequate depth, the sequence coverage of most proteins is far from complete and lags behind transcriptome analysis with RNA-seq. This implies that proteoforms[148,149] present due to alternative splicing are often not resolved in shotgun proteomics due to lack of sensitivity. Similarly, single-cell proteomics and plasma proteomics would substantially benefit from

improvements in sensitivity and dynamic range of measurements. Prediction of fragmentation spectra will help address these challenges by better integration of the intensity information into the available search engines. For this purpose, and also to accommodate PTMs, the intensity prediction models must be computationally efficient. Furthermore, the diversity of peptide classes, for instance due to labeling, PTMs and cross linking, that need to be considered, makes it seem unlikely that one big deep learning model that knows everything will be the preferred way to proceed. Instead, a multitude of specialized models, each one trainable with moderate effort and limited training data, should better accommodate the needs. Currently, it is an open question whether efficiency for models with less training data is best achieved by transfer learning or by reverting to simpler models without RNNs, such as wiNNer. Either way, spectral prediction will have an increasingly profound impact on data analysis in proteomics.

REFERENCES

1. Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
2. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews* (2013) doi:10.1021/cr3003533.
3. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
4. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).
5. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biol. Mass Spectrom.* **11**, 601 (1984).
6. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699–711 (2004).
7. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* (2018) doi:10.3390/metabo8020031.
8. Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biol. Mass Spectrom.* (1988) doi:10.1002/bms.1200160119.
9. Mitchell Wells, J. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* **402**, 148–185 (2005).
10. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).
11. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* (2004) doi:10.1073/pnas.0402700101.
12. Borges, R. M. *et al.* Quantum Chemistry Calculations for Metabolomics. *Chem. Rev.*

(2021) doi:10.1021/acs.chemrev.0c00901.
13. 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* **5**, 976–989 (1994).
14. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
15. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
16. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* (2004) doi:10.1021/ac049951b.
17. Boyd, R. & Somogyi, Á. The mobile proton hypothesis in fragmentation of protonated peptides: A perspective. *Journal of the American Society for Mass Spectrometry* vol. 21 1275–1278 (2010).
18. Tiwary, S. *et al.* High quality MS/MS spectrum prediction for data-dependent and -independent acquisition data analysis. *Nat Methods* (2019).
19. Verbruggen, S. *et al.* Spectral prediction features as a solution for the search space size problem in proteogenomics. *Mol. Cell. Proteomics* (2021) doi:10.1016/J.MCPRO.2021.100076.
20. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* (2021) doi:10.1038/s41467-021-23713-9.
21. Domokos, L., Hennberg, D. & Weimann, B. Computer-aided identification of compounds by comparison of mass spectra. *Anal. Chim. Acta* (1984) doi:10.1016/S0003-2670(00)85186-7.
22. Yates, J. R., Morgan, S. F., Gatlin, C. L., Griffin, P. R. & Eng, J. K. Method to Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis. *Anal. Chem.* (1998) doi:10.1021/ac980122y.
23. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* (1994) doi:10.1016/1044-0305(94)87009-8.
24. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* (2007) doi:10.1002/pmic.200600625.
25. Neuhauser, N., Michalski, A., Cox, J. & Mann, M. Expert system for computer-assisted annotation of MS/MS spectra. *Mol Cell Proteomics* **11**, 1500–1509 (2012).
26. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P. & Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* (2004) doi:10.1038/nbt930.
27. Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H. & Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput.* **230**, 219–230 (2006).
28. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
29. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning - An MIT Press book*. *MIT Press* (2016).
30. Zhou, X. X. *et al.* PDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **89**, 12690–12697 (2017).
31. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* (2019) doi:10.1038/s41592-019-0426-7.
32. Yang, Y., Lin, L. & Qiao, L. Deep learning approaches for data-independent acquisition proteomics. *Expert Rev. Proteomics* **18**, 1031–1043 (2021).
33. Wen, B. *et al.* Deep Learning in Proteomics. *Proteomics* (2020)

doi:10.1002/pmic.201900335.
34. Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Reports Methods* **1**, (2021).
35. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: A tutorial. *Molecular Systems Biology* (2008) doi:10.1038/msb.2008.61.
36. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* (2012) doi:10.1074/mcp.O111.016717.
37. Deutsch, E. W. *et al.* Expanding the Use of Spectral Libraries in Proteomics. *Journal of Proteome Research* (2018) doi:10.1021/acs.jproteome.8b00485.
38. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* (2004) doi:10.1038/nmeth705.
39. Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* (2013) doi:10.1038/nmeth.2528.
40. Distler, U. *et al.* Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* (2014) doi:10.1038/nmeth.2767.
41. Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial . *Mol. Syst. Biol.* (2018) doi:10.15252/msb.20178126.
42. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35–35 (2014).
43. Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* (1986) doi:10.1023/A:1022643204877.
44. Moore, D. H. Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984,358 pages, $27.95. *Cytometry* (1987) doi:10.1002/cyto.990080516.
45. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
46. Chen, T. & Guestrin, C. XGBoost : Reliable Large-scale Tree Boosting System. *arXiv* (2016) doi:10.1145/2939672.2939785.
47. Vapnik, V. N. *The nature of statistical learning theory*. (Springer, 1995).
48. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **9**, 155–161 (1997).
49. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* (1986) doi:10.1038/323533a0.
50. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation* (2019) doi:10.1162/neco_a_01199.
51. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* (1997) doi:10.1109/78.650093.
52. Hochreiter, S. & Schmidhuber, J. J. Long short-term memory. *Neural Comput.* **9**, 1–32 (1997).
53. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* (2000) doi:10.1162/089976600300015015.
54. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Gated feedback recurrent neural networks. in *32nd International Conference on Machine Learning, ICML 2015* (2015).
55. LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* (1989) doi:10.1162/neco.1989.1.4.541.
56. West, J., Ventura, D. & Warnick, S. Spring research presentation: A theoretical foundation for inductive transfer. *Brigham Young Univ. Coll. Phys. Math. Sci.* (2007).
57. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* (2017).

58. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *34th International Conference on Machine Learning, ICML 2017* (2017).
59. Marx, H. *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol* **31**, 557–564 (2013).
60. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).
61. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2020: Enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz984.
62. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkab1038.
63. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *EMBO Reports* (2008) doi:10.1038/embor.2008.56.
64. Wang, M. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.* (2018) doi:10.1016/j.cels.2018.08.004.
65. Okuda, S. *et al.* JPOSTrepo: An international standard data repository for proteomes. *Nucleic Acids Res.* (2017) doi:10.1093/nar/gkw1080.
66. Ma, J. *et al.* Iprox: An integrated proteome resource. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gky869.
67. Sharma, V. *et al.* Panorama public: A public repository for quantitative data sets processed in skyline. *Mol. Cell. Proteomics* (2018) doi:10.1074/mcp.RA117.000543.
68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
69. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
70. Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *J. Proteome Res.* (2008) doi:10.1021/pr070361e.
71. Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* (2016) doi:10.1038/nmeth.3902.
72. Savitski, M. M. *et al.* Targeted Data Acquisition for Improved Reproducibility and Robustness of Proteomic Mass Spectrometry Assays. *J. Am. Soc. Mass Spectrom.* **21**, 1668–1679 (2010).
73. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res* **10**, 1785–1793 (2011).
74. Wan, K. X., Vidavsky, I. & Gross, M. L. Comparing similar spectra: From similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.* (2002) doi:10.1016/S1044-0305(01)00327-0.
75. Liu, J. *et al.* Methods for peptide identification by spectral comparison. *Proteome Sci.* (2007) doi:10.1186/1477-5956-5-3.
76. Shao, W., Zhu, K. & Lam, H. Refining similarity scoring to enable decoy-free validation in spectral library searching. *Proteomics* (2013) doi:10.1002/pmic.201300232.
77. Garg, N. *et al.* Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures. *Int. J. Mass Spectrom.* (2015) doi:10.1016/j.ijms.2014.06.005.
78. Toprak, U. H. *et al.* Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol. Cell. Proteomics* (2014) doi:10.1074/mcp.O113.036475.
79. Li, S., Arnold, R. J., Tang, H. & Radivojac, P. On the accuracy and limits of peptide fragmentation spectrum prediction. *Anal. Chem.* (2011) doi:10.1021/ac102272r.

80. Tarn, C. & Zeng, W. F. PDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning. *Anal. Chem.* (2021) doi:10.1021/acs.analchem.0c05427.
81. Guan, S., Moran, M. F. & Ma, B. Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Mol. Cell. Proteomics* (2019) doi:10.1074/mcp.TIR119.001412.
82. Lin, Y. M., Chen, C. T. & Chang, J. M. MS2CNN: Predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks. *BMC Genomics* (2019) doi:10.1186/s12864-019-6297-6.
83. Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. in *Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (2014). doi:10.3115/v1/w14-4012.
84. Degroeve, S., Martens, L. & Jurisica, I. MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–3203 (2013).
85. Degroeve, S., Maddelein, D. & Martens, L. MS2PIP prediction server: Compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res.* (2015) doi:10.1093/nar/gkv542.
86. Gabriels, R., Martens, L. & Degroeve, S. Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz299.
87. Zhou, C., Bowler, L. D. & Feng, J. A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics* (2008) doi:10.1186/1471-2105-9-325.
88. Frank, A. M. Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* (2009) doi:10.1021/pr800677f.
89. Dong, N. P. *et al.* Prediction of Peptide Fragment Ion Mass Spectra by Data Mining Techniques. *Anal. Chem.* **86**, 7446–7454 (2014).
90. Welker, F. *et al.* The dental proteome of Homo antecessor. *Nature* (2020) doi:10.1038/s41586-020-2153-8.
91. Liu, K., Li, S., Wang, L., Ye, Y. & Tang, H. Full-Spectrum Prediction of Peptides Tandem Mass Spectra using Deep Neural Network. *Anal. Chem.* (2020) doi:10.1021/acs.analchem.9b04867.
92. Caruana, R. Multitask Learning. *Mach. Learn.* (1997) doi:10.1023/A:1007379606734.
93. French, R. M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences* (1999) doi:10.1016/S1364-6613(99)01294-2.
94. Frese, C. K. *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.* (2012) doi:10.1021/ac3025366.
95. Brodbelt, J. S., Morrison, L. J. & Santos, I. Ultraviolet Photodissociation Mass Spectrometry for Analysis of Biological Molecules. *Chemical Reviews* (2020) doi:10.1021/acs.chemrev.9b00440.
96. Zeng, W. F. *et al.* MS/MS Spectrum prediction for modified peptides using pDeep2 Trained by Transfer Learning. *Anal. Chem.* (2019) doi:10.1021/acs.analchem.9b01262.
97. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* (2021) doi:10.1038/s41592-021-01301-5.
98. Reily, C., Stewart, T. J., Renfrow, M. B. & Novak, J. Glycosylation in health and disease. *Nature Reviews Nephrology* (2019) doi:10.1038/s41581-019-0129-4.
99. Yang, Y., Horvatovich, P. & Qiao, L. Fragment Mass Spectrum Prediction Facilitates Site Localization of Phosphorylation. *J. Proteome Res.* (2021)

doi:10.1021/acs.jproteome.0c00580.
100. Lou, R. *et al.* DeepPhospho accelerates DIA phosphoproteome profiling through in silico library generation. *Nat. Commun.* (2021) doi:10.1038/s41467-021-26979-1.
101. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nature Structural and Molecular Biology* (2018) doi:10.1038/s41594-018-0147-0.
102. Chen, Z. L., Mao, P. Z., Zeng, W. F., Chi, H. & He, S. M. PDeepXL: MS/MS Spectrum Prediction for Cross-Linked Peptide Pairs by Deep Learning. *J. Proteome Res.* (2021) doi:10.1021/acs.jproteome.0c01004.
103. Giese, S. H., Sinn, L. R., Wegner, F. & Rappsilber, J. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nat. Commun.* (2021) doi:10.1038/s41467-021-23441-0.
104. Yılmaz, Ş., Busch, F., Nagaraj, N. & Cox, J. Accurate and automated high-coverage identification of chemically cross-linked peptides with MaxLynx. *bioRxiv* (2021).
105. Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* (2007) doi:10.1021/pr0604054.
106. Narasimhan, C. *et al.* MASPIC: Intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem.* (2005) doi:10.1021/ac0501745.
107. Sadygov, R., Wohlschlegel, J., Park, S. K., Xu, T. & Yates, J. R. Central limit theorem as an approximation for intensity-based scoring function. *Anal. Chem.* (2006) doi:10.1021/ac051206r.
108. Silva, A. S. C., Bouwmeester, R., Martens, L. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz383.
109. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkaa1100.
110. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
111. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* (2016) doi:10.1007/s13361-016-1460-7.
112. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* (2014) doi:10.1038/ncomms6277.
113. Chong, C., Coukos, G. & Bassani-Sternberg, M. Identification of tumor antigens with immunopeptidomics. *Nature Biotechnology* (2021) doi:10.1038/s41587-021-01038-8.
114. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114–1125 (2014).
115. Wilmes, P. & Bond, P. L. Metaproteomics: Studying functional gene expression in microbial ecosystems. *Trends in Microbiology* (2006) doi:10.1016/j.tim.2005.12.006.
116. Kloetzel, P. M. Antigen processing by the proteasome. *Nature Reviews Molecular Cell Biology* (2001) doi:10.1038/35056572.
117. Coulie, P. G. *et al.* A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. *Proc. Natl. Acad. Sci. U. S. A.* (1995) doi:10.1073/pnas.92.17.7976.
118. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* (2017) doi:10.1038/nature22991.

119. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* (2017) doi:10.1038/nature23003.
120. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science (80-.).* (1992) doi:10.1126/science.1546328.
121. Admon, A. & Bassani-Sternberg, M. The human immunopeptidome project, a suggestion for yet another postgenome next big thing. *Molecular and Cellular Proteomics* (2011) doi:10.1074/mcp.O111.011833.
122. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* (2020) doi:10.1002/pmic.201900334.
123. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-019-0322-9.
124. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
125. Sinitcyn, P. *et al.* MaxQuant goes Linux. *Nat. Methods* **15**, 401 (2018).
126. Liepe, J. *et al.* A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science (80-.).* (2016) doi:10.1126/science.aaf4384.
127. Faridi, P. *et al.* A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Sci. Immunol.* (2018) doi:10.1126/sciimmunol.aar3947.
128. Specht, G. *et al.* Large database for the analysis and prediction of spliced and non-spliced peptide generation by proteasomes. *Sci. Data* (2020) doi:10.1038/s41597-020-0487-6.
129. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* (2017) doi:10.1016/j.ymeth.2017.05.028.
130. Garalde, D. R. *et al.* Highly parallel direct RN A sequencing on an array of nanopores. *Nat. Methods* (2018) doi:10.1038/nmeth.4577.
131. Schoenholz, S. S. *et al.* Peptide-Spectra Matching from Weak Supervision. *arXiv Prepr. arXiv1808.06576* (2018).
132. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
133. Li, Y. *et al.* Group-DIA: Analyzing multiple data-independent acquisition mass spectrometry data files. *Nature Methods* (2015) doi:10.1038/nmeth.3593.
134. Bekker-Jensen, D. B. *et al.* Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.* (2020) doi:10.1038/s41467-020-14609-1.
135. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
136. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology* (2014) doi:10.1038/nbt.2841.
137. Bruderer, R. *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* (2015) doi:10.1074/mcp.M114.044305.
138. Keller, A., Bader, S. L., Shteynberg, D., Hood, L. & Moritz, R. L. Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition mass spectrometry (MS) using SWATHProphet. *Mol. Cell. Proteomics* (2015) doi:10.1074/mcp.O114.044917.
139. Meyer, J. G. *et al.* PIQED: Automated identification and quantification of protein

modifications from DIA-MS data. *Nature Methods* (2017) doi:10.1038/nmeth.4334.
140. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* (2018) doi:10.1038/s41467-018-07454-w.
141. Peckner, R. *et al.* Specter: Linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nat. Methods* (2018) doi:10.1038/nmeth.4643.
142. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* (2020) doi:10.1038/s41592-019-0638-x.
143. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00968-7.
144. Searle, B. C. *et al.* Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* (2020) doi:10.1038/s41467-020-15346-1.
145. Lou, R. *et al.* Hybrid Spectral Library Combining DIA-MS Data and a Targeted Virtual Library Substantially Deepens the Proteome Coverage. *iScience* (2020) doi:10.1016/j.isci.2020.100903.
146. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* (2020) doi:10.1038/s41467-019-13866-z.
147. Isaksson, M., Karlsson, C., Laurell, T., Kirkeby, A. & Heusel, M. MSLibrarian: Optimized Predicted Spectral Libraries for Data-Independent Acquisition Proteomics. *J. Proteome Res.* **21**, 535–546 (2022).
148. Smith, L. M. & Kelleher, N. L. Proteoforms as the next proteomics currency. *Science (80-.).* **359**, 1106–1107 (2018).
149. Aebersold, R. *et al.* How many human proteoforms are there? *Nature Chemical Biology* vol. 14 206–214 (2018).
150. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science (80-.).* **246**, 64–71 (1989).
151. Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. *Anal. Chem.* **63**, 1193A-1203A (1991).
152. Bateman, R. H. *et al.* A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. *J. Am. Soc. Mass Spectrom.* (2002) doi:10.1016/S1044-0305(02)00420-8.
153. Geiger, T., Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics* **9**, 2252–2261 (2010).
154. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A Neural Probabilistic Language Model. in *Journal of Machine Learning Research* (2003). doi:10.1162/153244303322533223.
155. Coscia, F. *et al.* A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. *J. Pathol.* (2020) doi:10.1002/path.5420.

ACKNOWLEDGEMENTS

I thank G. Borner, B. Frohn, T. Geiger, J. L. Restrepo-López, F. Traube and S. Yilmaz for critical reading of the manuscript, C. De Nart for assistance with the figure in Box 1 and P. Sinitcyn for the re-analysis of data in Fig. 6a. This project was partially funded by the German Ministry for Science and Education (BMBF) funding action MSCoreSys, reference number FKZ 031L0214D.

COMPETING FINANCIAL INTERESTS

The author declares no competing interests.

FIGURE LEGENDS

Fig. 1 | Fragmentation spectra in shotgun proteomics. a, The shotgun proteomics workflow starts by extracting proteins from samples of interest and digesting them to peptides by a protease, most often trypsin. Peptides are then separated by liquid chromatography (LC) and ionized[150,151]. First-level MS or MS1 spectra record the mass to charge ratios of peptides. In data-dependent acquisition (DDA), peptide precursors are selected in narrow isolation windows aiming at selecting single molecular species and subjected to fragmentation. The resulting fragmentation spectra, also called MS/MS or MS2, contain the masses of the resulting fragments which are dominated by characteristic series of ions. **b,** A peptide is a chain of amino acids of arbitrary length. The example shows a peptide of length four, in which the residues R_1 to R_4 can be any of the 20 standard amino acid side chains. The blue symbols indicate products of main chain bond breakages which, if charged, can be detected in the mass spectrometer. **c,** A typical fragmentation spectrum of a peptide without modifications and obtained by collisional dissociation, is dominated by *y*- and *b*-ion series (blue). These correspond to fragments resulting from a single bond breakage of a peptide bond in the main chain. The b-series is generated by the N-terminal and the y-series by the C-terminal pieces. Additional non-regular fragments can be generated by neutral losses of H_2O or NH_3 molecules (green). In principle, the spectrum can contain other types of fragments originating from the precursor peptide, such as less common neutral losses or internal fragments from two simultaneous main chain bond breakages (orange), but these are usually less frequent and low abundant. However, these nonstandard fragments are in principle predictable and therefore useful for peptide identification. Furthermore, other peptides that have the same or a similar precursor m/z and retention time can be involuntarily co-fragmented resulting in peaks that cannot be accounted for by the peptide of interest (red). Additional complexity arises from the possibility that fragments can carry more than one positive charge, in particular for precursors of charge three or higher. Other fragmentation methods create spectra that are dominated by other types of ion series due to different bond breakages in the main chain. For instance, spectra created by electron transfer dissociation are dominated by *c*- and *z*-type ions.

Fig. 2 | DDA and DIA. A mass spectrometric cycle typically consists of a full scan recording the signals of peptides that are currently eluting from the liquid chromatography plus a number of MS/MS spectra, three in this example, containing fragmentation signals. **a,** In

DDA mode ions are isolated for fragmentation in narrow windows that change from cycle to cycle. **b,** In DIA mode the ions are selected in sets of wide mass windows that are the same in all cycles. The window sizes can range from a few Dalton[41] to a single window covering the whole mass range[152,153].

Fig. 3 | Machine learning. a, The fragmentation spectrum prediction is a regression problem with peptide sequences and metadata as input and spectral intensities as output. The sequence is usually fed in by either one-hot encoding, corresponding to indicator variables, or through an embedding layer[154]. One-hot encoding uses binary vectors of length 20 per sequence position, containing only zeroes and ones as entries encoding the standard amino acids, and possibly more to represent modified amino acids. An embedding layer finds linear combinations of input features, akin to a principal component analysis that best represent them. One-hot encoding is just a particular choice of variables, whereas an embedding layer dynamically determines the feature representation and adds to the computation time for model training. **b,** A computational neuron receives the numerical inputs a_j which are modulated by excitatory (positive) or inhibitory (negative) weights w_j, biased by w_0 and evaluated by the activation function g to produce the output which is fed into other computational neurons. **c,** A multilayer perceptron arranges computational neurons in a multi-layered structure with a directional ('feed-forward') information flow. **d,** Transfer learning. Part of the model trained on big data is frozen and only the remaining layers are trained on a smaller specialized dataset. **e,** The available data is split into three parts, training, test and validation set, which serve different purposes. **f,** Cross validation. The example shows five-fold cross validation, in which the available data is split into five equally sized parts. Each part serves once as the test set while the remaining data comprises the training and validation parts. **g,** Typical dependence of prediction accuracy on the number of training instances. The more training examples are used, the better the correlation between true and predicted spectra. Both models reach some asymptotic performance at an infinite number of training instances. Although model B's asymptotic performance is worse in the example it learns better if only little training data is available.

Fig. 4 | Machine learning strategies for ion series intensity prediction. a, Bidirectional RNN architecture of DeepMass:Prism. **b,** Fixed peptide length models, e.g. implemented in

MS2PIP, which uses random forests. For ease of comparison a multilayer perceptron is used here as well. **c,** Sequence window-based prediction with wiNNer.

Fig. 5 | DDA applications. a, Identification rate improvement as a function of q-value on a HeLa dataset when integrating intensity predictions by DeepMass:Prism, wiNNer and MS2PIP into the Andromeda score. (Adapted from Fig. 6b in ref. [18].) **b,** Peptides gained, shared, and lost when rescoring MaxQuant results with intensity prediction information compared with Spectrum Mill analysis[123] on 92 monoallelic cell lines. (Adapted from Fig. 3a in ref. [20]) **c,** The number of identified spectra as a function of varying FDR levels for RNA-seq based proteogenomics searches with (orange) and without (blue) using spectral prediction information. (Adapted from Fig. 1b in ref. [19].)

Fig. 6 | DIA applications. a, Principal component analysis of tissue samples measured with DIA runs. Ellipses indicate samples of same tissue origin. (Adapted from Fig. 2f in ref. [155].) **b,** Re-analysis of the data in a, using predicted libraries in MaxDIA. **c,** Venn diagram of the number of genes covered by protein groups in the analysis of fractionated HEK cell lysate when using three different libraries. (Based on Fig. 6c in ref. [143].) **d,** Venn diagrams of gene and peptide counts when using three different library prediction methods (Adapted from Supplementary Fig 11 in ref. [143].)

FIGURES

Figure 1

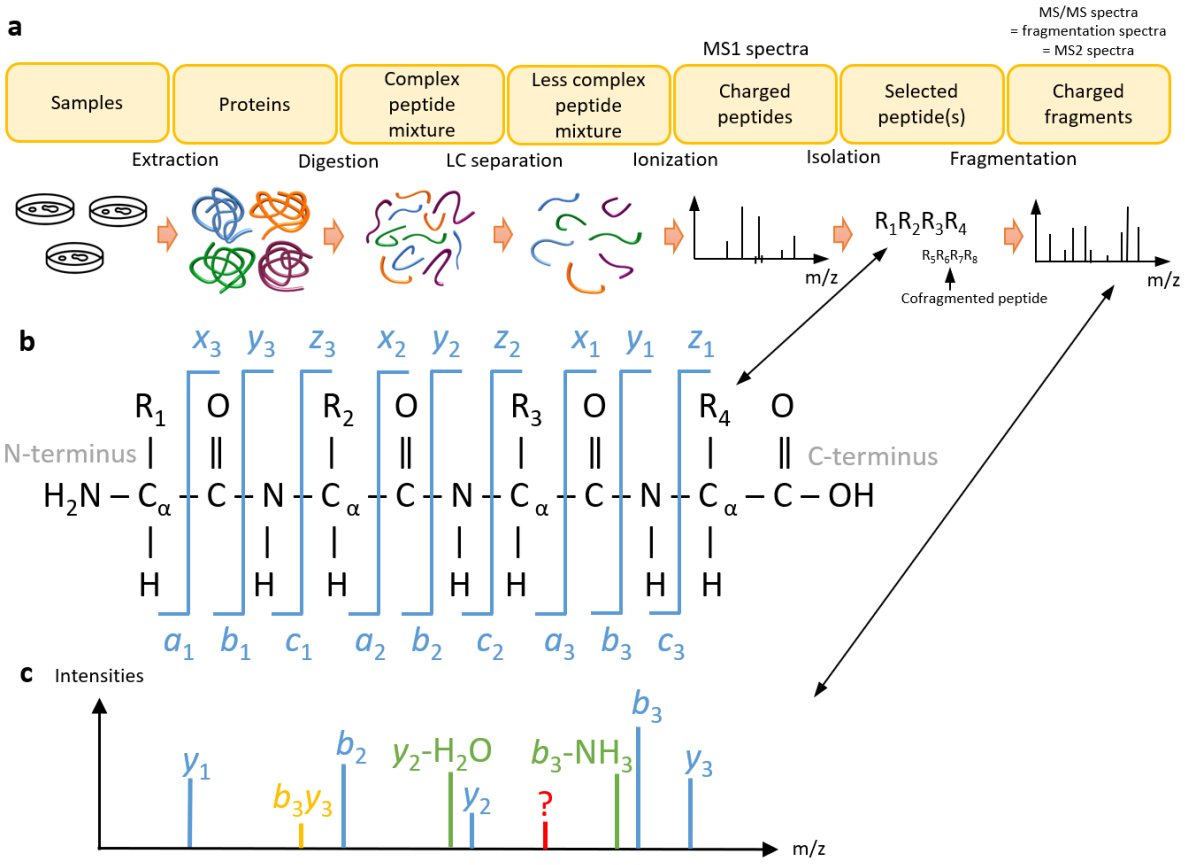


Figure 2

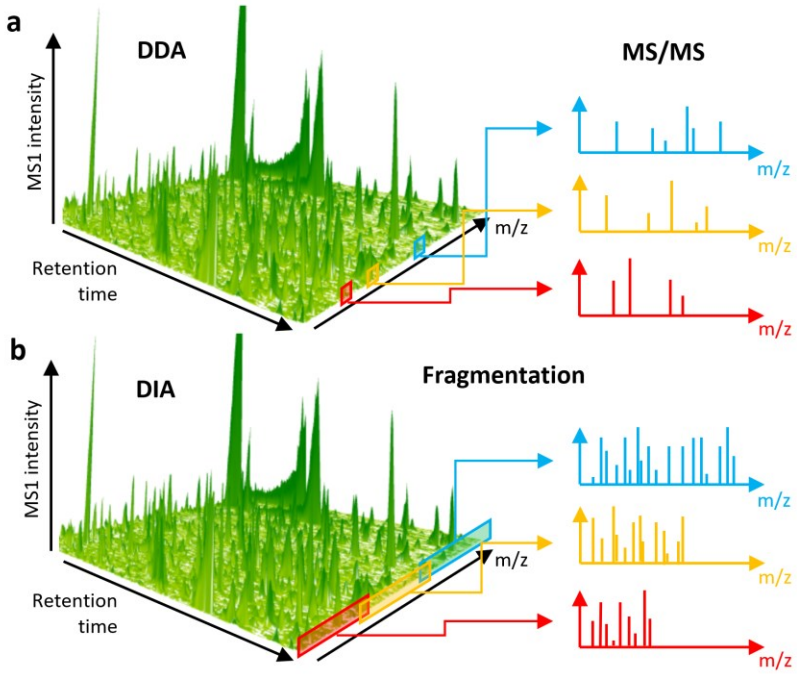

a
DDA
MS/MS
MS1 intensity
Retention time
m/z
m/z
m/z
m/z
b
DIA
Fragmentation
MS1 intensity
Retention time
m/z
m/z
m/z
m/z

Figure 3

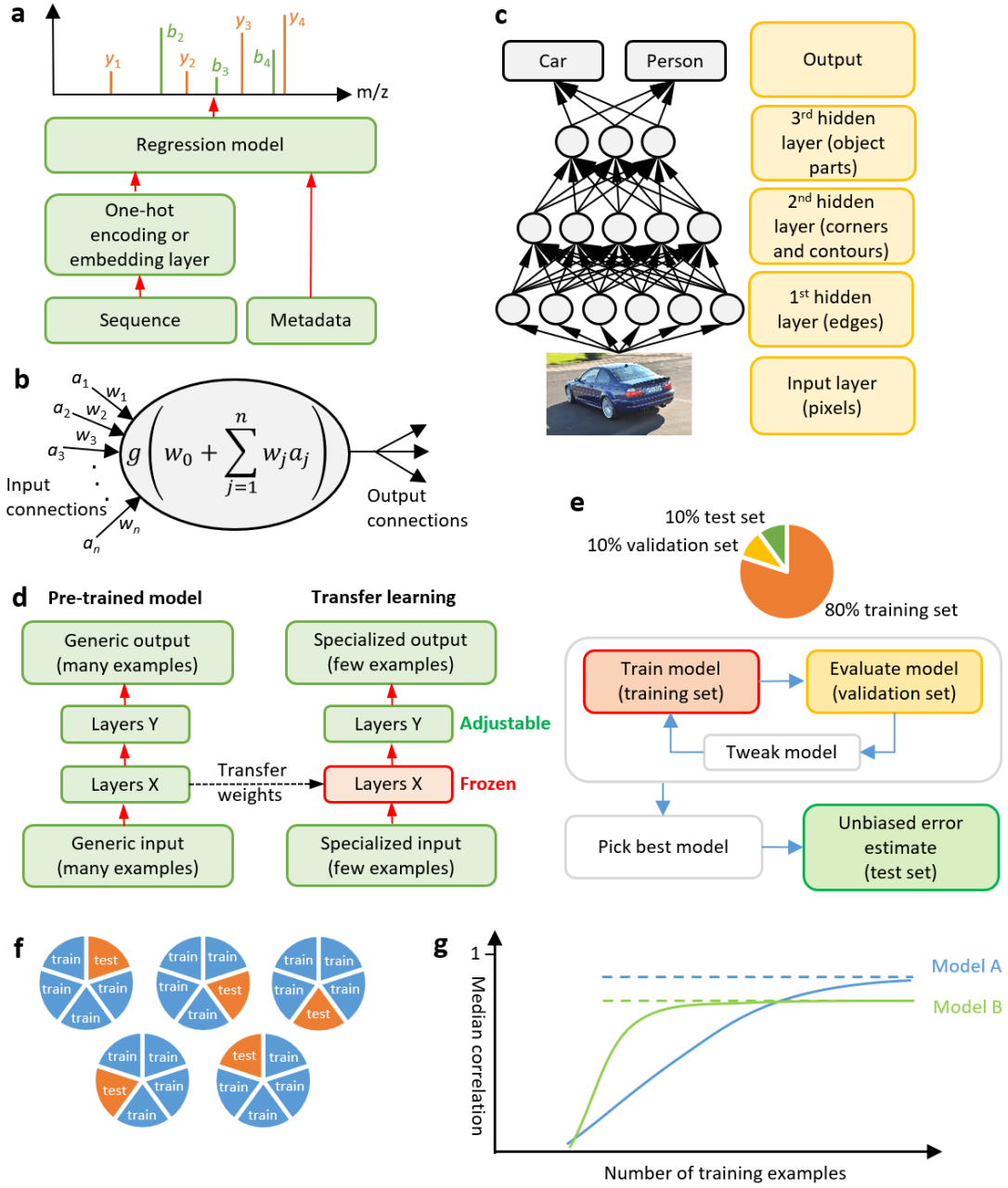

a
m/z
Regression model
One-hot encoding or embedding layer
Sequence
Metadata
b
Input connections
Output connections
c
Car
Person
Output
3rd hidden layer (object parts)
2nd hidden layer (corners and contours)
1st hidden layer (edges)
Input layer (pixels)
d
Pre-trained model
Transfer learning
Generic output (many examples)
Layers Y
Layers X
Generic input (many examples)
Transfer weights
Specialized output (few examples)
Adjustable
Frozen
Specialized input (few examples)
e
10% test set
10% validation set
80% training set
Train model (training set)
Evaluate model (validation set)
Tweak model
Pick best model
Unbiased error estimate (test set)
f
train
test
g
Median correlation
Model A
Model B
Number of training examples

Figure 4

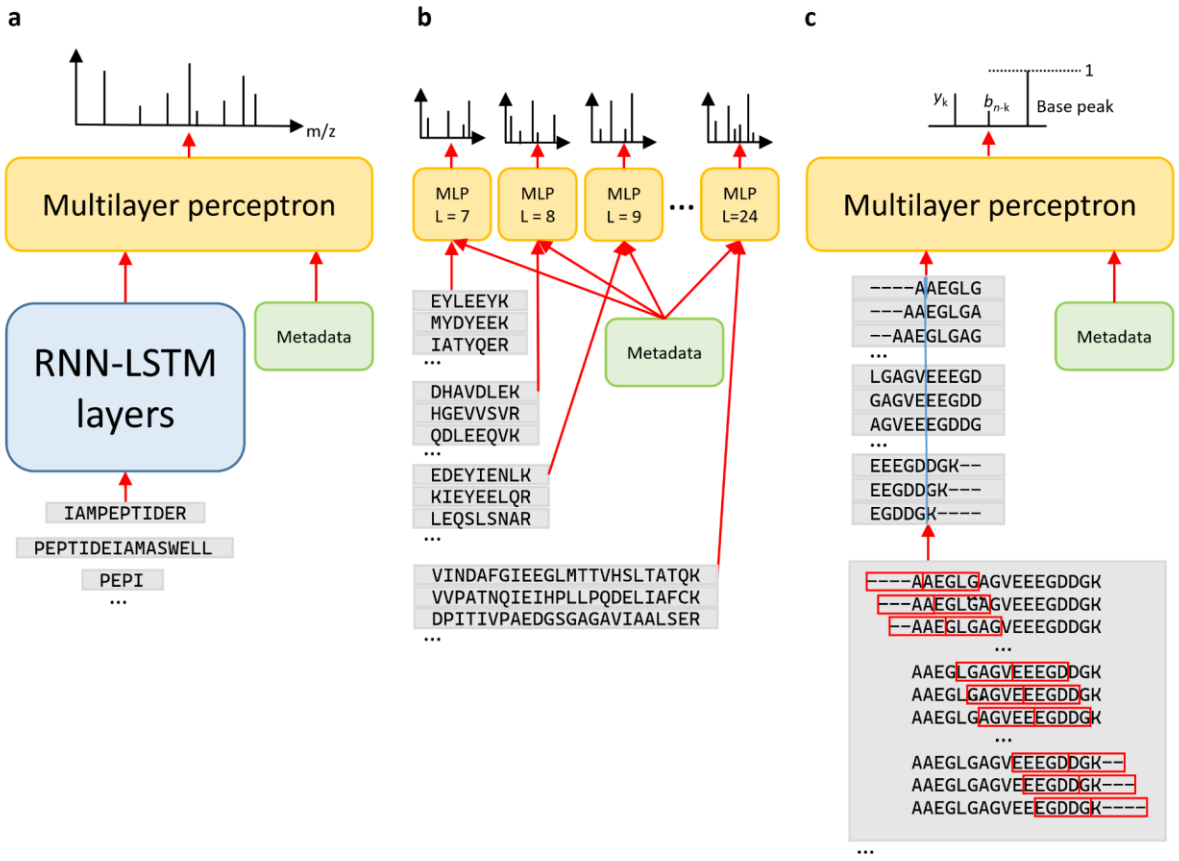

a
m/z
Multilayer perceptron
RNN-LSTM layers
Metadata
IAMPEPTIDER
PEPTIDEIAMASWELL
PEPI
...
b
MLP L = 7
MLP L = 8
MLP L = 9
MLP L=24
Metadata
EYLEEYK
MYDYEEK
IATYQER
DHAVDLEK
HGEVVSVR
QDLEEQVK
EDEYIENLK
KIEYEELQR
LEQSLSNAR
VINDAFGIEEGLMTTVHSLTATQK
VVPATNQIEIHPLLPQDELIAFCK
DPITIVPAEDGSGAGAVIAALSER
c
1
y_k
b_{n-k}
Base peak
Multilayer perceptron
Metadata
----AAEGLG
---AAEGLGA
--AAEGLGAG
LGAGVEEEGD
GAGVEEEGDD
AGVEEEGDDG
EEEGDDGK--
EEGDDGK---
EGDDGK----
----AAEGLGAGVEEEGDDGK
---AAEGLGAGVEEEGDDGK
--AAEGLGAGVEEEGDDGK
AAEGLGAGVEEEGDDGK
AAEGLGAGVEEEGDDGK
AAEGLGAGVEEEGDDGK
AAEGLGAGVEEEGDDGK--
AAEGLGAGVEEEGDDGK---
AAEGLGAGVEEEGDDGK----

Figure 5

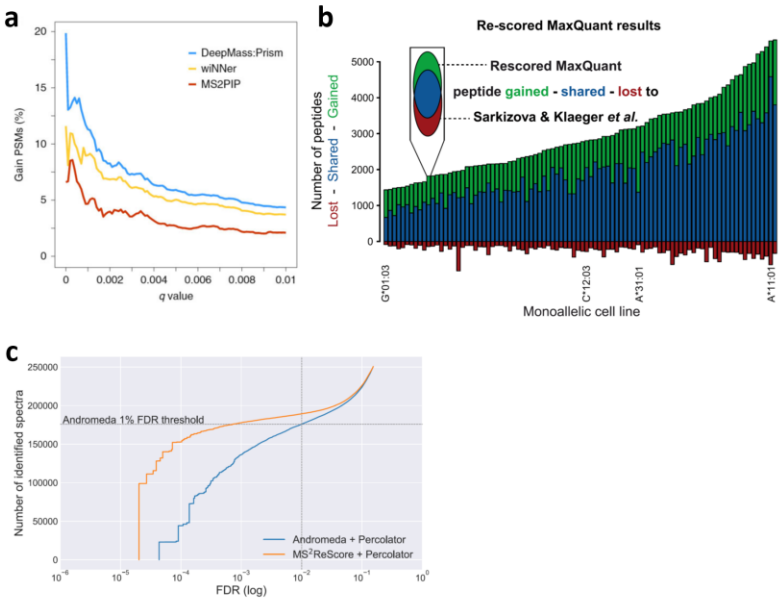
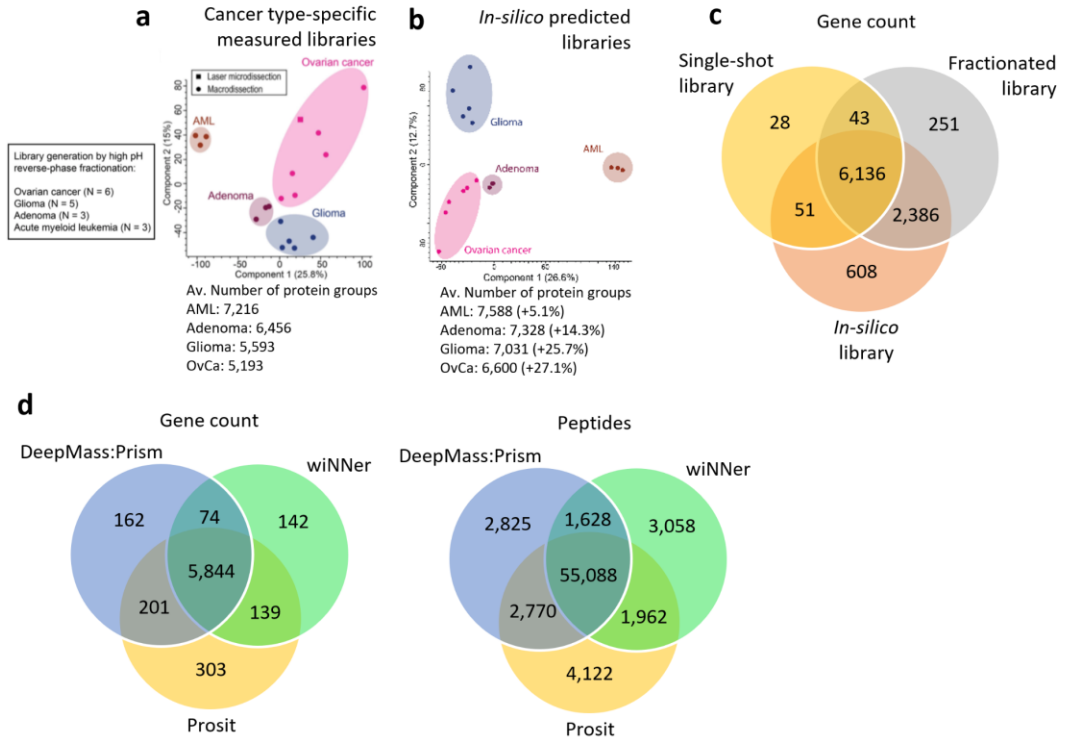

a
DeepMass:Prism
wiNNer
MS2PIP
Gain PSMs (%)
q value
b
Re-scored MaxQuant results
Rescored MaxQuant
peptide gained - shared - lost to
Sarkizova & Klaeger et al.
Number of peptides
Lost - Shared - Gained
Monoallelic cell line
c
Andromeda 1% FDR threshold
Number of identified spectra
Andromeda + Percolator
MS²ReScore + Percolator
FDR (log)

Figure 6


a
Cancer type-specific measured libraries
Library generation by high pH reverse-phase fractionation:
Ovarian cancer (N = 6)
Glioma (N = 5)
Adenoma (N = 3)
Acute myeloid leukemia (N = 3)
Laser microdissection
Macrodissection
Ovarian cancer
AML
Adenoma
Glioma
Component 1 (25.8%)
Component 2 (15%)
Av. Number of protein groups
AML: 7,216
Adenoma: 6,456
Glioma: 5,593
OvCa: 5,193
b
In-silico predicted libraries
Glioma
AML
Adenoma
Ovarian cancer
Component 1 (26.6%)
Component 2 (12.7%)
Av. Number of protein groups
AML: 7,588 (+5.1%)
Adenoma: 7,328 (+14.3%)
Glioma: 7,031 (+25.7%)
OvCa: 6,600 (+27.1%)
c
Gene count
Single-shot library
Fractionated library
28
43
251
6,136
51
2,386
608
In-silico library
d
Gene count
DeepMass:Prism
wiNNer
162
74
142
5,844
201
139
303
Prosit
Peptides
DeepMass:Prism
wiNNer
2,825
1,628
3,058
55,088
2,770
1,962
4,122
Prosit