nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0426-7

Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning

Siegfried Gessulat[1,2,7], Tobias Schmidt[1,7], Daniel Paul Zolg[1], Patroklos Samaras[1], Karsten Schnatbaum[3], Johannes Zerweck[3], Tobias Knaute[3], Julia Rechenberger[1], Bernard Delanghe[4], Andreas Huhmer[5], Ulf Reimer[3], Hans-Christian Ehrlich[2], Stephan Aiche[2], Bernhard Kuster[1,6*] and Mathias Wilhelm[1*]

In mass-spectrometry-based proteomics, the identification and quantification of peptides and proteins heavily rely on sequence database searching or spectral library matching. The lack of accurate predictive models for fragment ion intensities impairs the realization of the full potential of these approaches. Here, we extended the ProteomeTools synthetic peptide library to 550,000 tryptic peptides and 21 million high-quality tandem mass spectra. We trained a deep neural network, termed Prosit, resulting in chromatographic retention time and fragment ion intensity predictions that exceed the quality of the experimental data. Integrating Prosit into database search pipelines led to more identifications at >10× lower false discovery rates. We show the general applicability of Prosit by predicting spectra for proteases other than trypsin, generating spectral libraries for data-independent acquisition and improving the analysis of metaproteomes. Prosit is integrated into ProteomicsDB, allowing search result re-scoring and custom spectral library generation for any organism on the basis of peptide sequence alone.

Mass spectrometry and computational data analysis have transformed proteomics research and enable proteome-scale analysis of biological systems[1]. In the prevalent bottom-up approach, proteins are digested by a protease and the resulting peptides are analyzed by liquid chromatography–tandem mass spectrometry (LC–MS/MS)[2]. Matching fragment ion spectra to peptide sequences is at the heart of peptide (and by inference protein) identification, quantification and all subsequent biological interpretation[3]. The de facto standard approach today is database searching[4], in which a fragmentation spectrum is matched to theoretical spectra for candidate peptides generated in silico. Most commonly used search engines[5–7] score peptide spectrum matches (PSMs) on the presence of fragment ions but largely disregard fragment ion intensities or information regarding which fragment ions may be experimentally observed. Spectral library searching is a complementary approach[8,9], in which intensities of fragment ions from experimental spectra are correlated to library spectra, typically constructed from previous peptide identification data[10]. However, so far, spectral libraries are most commonly used for the analysis of targeted or data-independent acquisition (DIA) experiments[11–13]. In DIA, additional information, such as peptide retention time, is crucial to ensure confident peptide identification and quantification[14]. While many computational models have been developed for the prediction of retention times[15,16] and attempts have been made for predicting fragment ion intensities[17–19], models that accurately predict the latter are only recently emerging[20,21]. However, these models are often specific to their training data and need to be re-trained for specific laboratory conditions to deliver high-quality results, which limits their scope and applicability.

Taking advantage of the very large number of synthetic peptides and tandem mass spectra thereof generated within the ProteomeTools project[22], we report a deep learning architecture termed Prosit. Prosit can learn and predict both the chromatographic retention time and the fragment ion intensity of any peptide with extremely high quality. We demonstrate the merits of Prosit on a number of challenging examples and provide the scientific community with ready-to-use tools.

Results

Accurate retention time and fragment ion intensity prediction by deep learning. The ProteomeTools project[23] aims to provide high-quality reference MS/MS data of synthetic peptides, covering the entire human proteome and important post-translational modifications. Here we extend the publicly available data by 219,125 peptides (Fig. 1a and Supplementary Table 1), resulting in 8,417,953 additional high-quality (Andromeda score > 100) reference spectra covering five different fragmentation techniques and six different collision energies (Methods). We also systematically recorded chromatographic retention indices (iRT)[22]. ProteomeTools now contains 21,764,501 high-quality spectra from 576,256 unique precursors (peptide sequences, modifications and charge states). These cover 19,749 of the 20,040 human protein coding genes (98.5%; Supplementary Fig. 1) and distinguish 26,549 of the 42,164 Swiss-Prot annotated isoforms (63%).

Using this resource, we developed a generic artificial neural network architecture (Methods), termed Prosit (Latin for 'of benefit'), that enables high-accuracy predictions for retention time and fragment ion intensities. Prosit was inspired by neural machine translation[24] and uses recent advances in deep learning[25,26]. Because the model is split into an encoder and a decoder (Fig. 1b and Supplementary Fig. 2a), manual feature engineering can be avoided, allowing for the embedding of, for example, peptide sequences and additional parameters into an intermediate latent space. Decoders then use the learned representation to make predictions

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany. [2]SAP SE, Potsdam, Germany. [3]JPT Peptide Technologies GmbH, Berlin, Germany. [4]Thermo Fisher Scientific, Bremen, Germany. [5]Thermo Fisher Scientific, San Jose, CA, USA. [6]Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany. [7]These authors contributed equally: Siegfried Gessulat, Tobias Schmidt. *e-mail: kuster@tum.de; mathias.wilhelm@tum.de

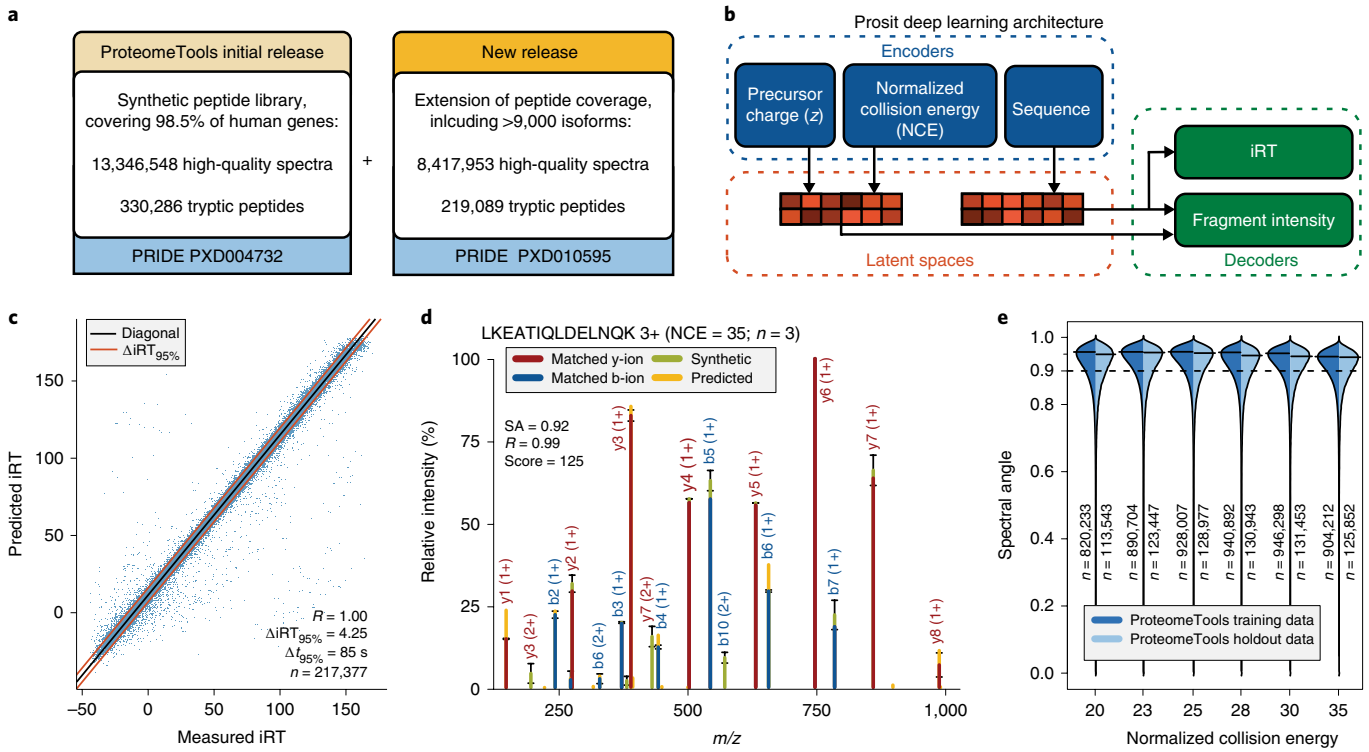

Fig. 1 | Accurate retention time and fragment ion intensity prediction by deep learning. a, The initial ProteomeTools dataset (right) of high-quality synthetic peptide spectra was extended in this study by 219,089 tryptic peptides extending the coverage of human proteins and isoforms. All data are available in ProteomeXchange under the accessions provided in the figure. **b**, The Prosit deep learning architecture for indexed retention time (iRT) and fragment ion intensity prediction. The input data (peptide sequences, precursor charge state and normalized collision energy) are encoded into a latent representation (space). This representation is then decoded to predict iRT (using sequences only) and fragment ion intensities (using all input parameters). **c**, Comparison of the measured versus predicted iRT values for 217,377 peptides not contained in the training data showed very high agreement (Pearson $R = 1.00$). Solid red lines mark the iRT isolation window required to encompass 95% of all peptides ($\Delta iRT_{95\%}$, $\Delta t_{95\%}$) around the diagonal (blue line). **d**, Pseudo mirror plot of a tandem mass spectrum comparing predicted fragment ion intensities for the triply charged peptide LKEATIQLDELNQK at an NCE of 35 to the spectrum of the synthetic peptide ($n = 3$ spectra, black error bars indicate 1 s.d. around the measured fragment ion intensities and the color change between error bars the median). Red and blue portions of each signal show the overlap between predicted and observed intensity for y and b ions, respectively. Green portions indicate that intensities of the spectrum of the synthetic peptide exceed the predicted intensities. Yellow portions indicate that predicted intensities exceed the intensities of the spectrum of the synthetic peptide. The Pearson correlation (R), normalized spectral contrast angle (SA) and Andromeda score (Score) are indicated. **e**, Distributions of spectral angles comparing measured and predicted tandem mass spectra at different NCEs for peptides contained in the training (dark blue) or holdout (light blue) set. Individual black bars indicate the apex of each distribution, and the dashed black line is drawn at a spectral angle of 0.9 ($R \sim 0.99$).

(Supplementary Notes). The data were split into three parts: one for training the model, a test set to control for overfitting and a holdout set to estimate performance (Supplementary Notes). Unless otherwise stated, performance is reported for the holdout set.

We initially trained Prosit to predict retention time indices (iRT). The model achieved very high agreement between predicted and measured values ($R = 1.00$; Fig. 1c) where 95% of the observations ($\Delta iRT_{95\%}$) were within 4.25 iRT units, corresponding to 85 s ($\Delta t_{95\%}$) for a 1-h LC–MS gradient. Prosit outperformed SSRCalc[27] on the same data ($R = 0.96$, $\Delta t_{95\%}$ of 20.4 iRT units; Supplementary Fig. 3 and Supplementary Notes). Next, we tested Prosit on the much more complex task of fragment ion intensity prediction using higher energy collision-induced dissociation (HCD) spectra of unmodified and methionine-sulfoxide-containing peptides (b and y ions only; Methods). In addition to the peptide sequence, the measured precursor ion charge state and the applied normalized collision energy (NCE) were used as inputs for training. To avoid having to learn models for each combination of parameters, we trained a single model by adding a second encoder that further modulates the latent representation (Supplementary Fig. 2a). Figure 1d shows a representative error plot (pseudo mirror plot) comparing the experimental and predicted spectra of a peptide of median prediction quality (Pearson $R = 0.99$), illustrating the exceptionally strong agreement between prediction and observation. Prosit uses normalized spectral angle as an objective function because of its higher sensitivity as a similarity measure (Supplementary Fig. 2c and Supplementary Notes)[28]. Predictions achieved very high agreement (median spectral angle of 0.92; median $R = 0.99$) across all investigated NCEs (Fig. 1e) without substantial overfitting (Supplementary Fig. 2b). Prosit also outperformed two recently reported fragment ion intensity prediction models (MS2PIP[20] and pDeep[20,21]) in virtually all cases (Supplementary Fig. 4 and Supplementary Notes).

Collision energy calibration improves fragment intensity prediction. The appearance of tandem mass spectrometry spectra strongly depends on the collision energy used for acquisition (Supplementary Fig. 5)[29]. Figure 2a shows that although Prosit made accurate predictions at a given NCE, a change in NCE had a dramatic effect on both fragment ion intensities and occurrence. Because Prosit was trained on six NCEs (Methods), we hypothesized that Prosit may be able to interpolate between NCEs using its NCE encoder. To test this, we compared Prosit predictions to data acquired for 15 different

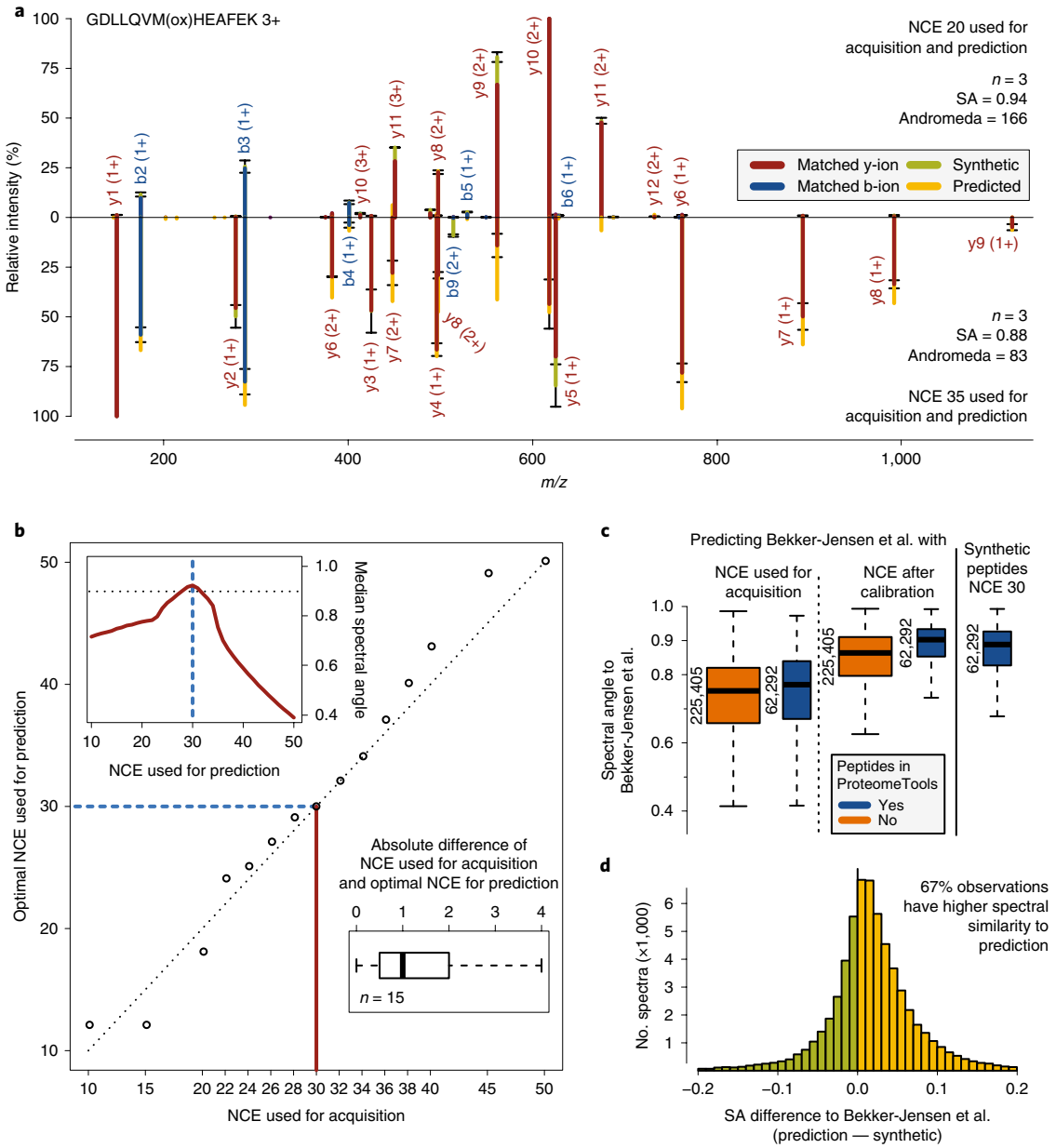

Fig. 2 | Collision energy calibration yields fragment intensity predictions with near-synthetic peptide spectrum quality. a, Double pseudo mirror plot of the synthetic, triply charged peptide GDLLQVM(ox)HEAFEK. The top panel compares the measured ($n = 3$; black error bars indicate 1 s.d. around the measured fragment ion intensities and the color change between error bars the median) and predicted spectra at NCE 20. The bottom panel shows the same analysis for NCE 35. Red and blue portions of each signal show the overlap between predicted and observed intensity for y and b ions, respectively. Green portions indicate that intensities of the spectrum of the synthetic peptide exceed the predicted intensities. Yellow portions indicate that predicted intensities exceed the intensities of the spectrum of the synthetic peptide. **b**, The top left inset shows the median spectral angle of comparing the spectra of 40 synthetic peptides measured at NCE30 to spectra predicted at NCEs of 10–50 (in steps of 1). Optimal agreement is reached at NCE30 (dashed blue line). The results of repeating this analysis for 15 different NCEs (open circles) used for data acquisition are shown in the large plot. The bottom right inset shows a box plot of the absolute differences between the NCE used for acquisition and the NCE at which predictions and measurements showed the strongest agreement. The box indicates the interquartile range (IQR), its whiskers 1.5× IQR values, and the black line the median; no outliers are shown. **c**, The left two box plots show the distribution of spectral angle (SA) values between spectra measured by Bekker-Jensen et al. versus spectra predicted by Prosit at the same collision energy. Peptides for which spectra are available in ProteomeTools and the Bekker-Jensen data are shown in blue; peptides present only in the work from Bekker-Jensen et al. are shown in orange. Measurement and prediction are in reasonably good agreement (SA, ~0.7; Pearson *R*, ~0.9). When Prosit was allowed to decide at which NCE the agreement between measurement and prediction was maximized (**b**), spectral angle values increased substantially (middle two box plots) and were very similar to spectral angle values obtained from directly comparing spectra of peptides in the work by Bekker-Jensen et al. with the same peptides in the ProteomeTools peptide set at the same collision energy. The boxes indicate the IQR and its whiskers indicate 1.5× IQR values; no outliers are shown. The median spectral angle and number of peptides are indicated. **d**, Histogram of the difference between SAs resulting from the comparison of predicted versus Bekker-Jensen spectra and synthetic peptide versus Bekker-Jensen spectra. On average, predicted spectra agree better with the Bekker-Jensen spectra (yellow) than the synthetic peptide spectra (green).

NCEs[22] (Fig. 2b). Comparing spectra predicted at every NCE between 10 and 50 (in steps of 1) to the acquired spectra generated a bell-shaped calibration curve (Fig. 2b, top inset) whose apex represents the NCE at which the predictions match best to the acquired spectra. Applying this across all 15 NCEs showed that Prosit very closely calibrates its prediction to the NCE used for acquisition with an absolute median NCE offset of 1 (Fig. 2b, bottom inset).

We next applied Prosit's ability to generalize to different NCEs to the analysis of external datasets without re-training the model. Figure 2c shows that the median spectral angle increased from 0.78 to 0.89 for a large external dataset (the Bekker-Jensen et al. dataset[30]) when we used the original or calibrated NCE, respectively. After calibration, the overall quality of Prosit predictions (median spectral angle, 0.913) was slightly higher compared to the experimental reference data from synthetic ProteomeTools peptides (median spectral angle, 0.907; Fig. 2c,d), indicating that the interpolation between collision energies worked very well. While peptides not synthesized in the ProteomeTools project showed a slightly lower overall spectral angle as a result of an overrepresentation of higher precursor charge states (Supplementary Fig. 6a), the overall prediction accuracy was consistent with data from the ProteomeTools project (Supplementary Fig. 6b) and outperformed other models (Prosit: spectral angle, 0.89, $R=0.99$; MSPIP: spectral angle, 0.65, $R=0.88$; Supplementary Fig. 6c, Supplementary Table 2 and Supplementary Notes).

Prosit predictions generalize to non-tryptic peptides. As shown above for tryptic peptides, Prosit can be calibrated on NCEs in the same fashion on non-tryptic peptides (Fig. 3a) without training a new model. LysC showed higher overall correlation (spectral angle, 0.88) than chymotrypsin (spectral angle, 0.86) and GluC (spectral angle, 0.82), likely because of its overlapping substrate specificity with trypsin (Fig. 3b). These results suggest that Prosit has learned substantial general characteristics of peptide fragmentation, but also that it can be improved further by the inclusion of non-tryptic peptides during training.

Plotting predicted and observed retention times for tryptic or chymotryptic peptides showed that the retention time for both classes could be predicted with high accuracy ($R=0.89$ and $R=0.91$, respectively; Fig. 3c, left). However, for both proteases, a group of outlier peptides with high predicted but low observed retention times were detected, indicating that additional liquid chromatography parameters not captured by iRT values led to suboptimal predictions by Prosit's basic model (Supplementary Table 3 and Supplementary Notes). We therefore applied transfer learning (Methods) by initializing Prosit using the ProteomeTools iRT model and further trained it on the basis of the tryptic peptides identified by Bekker-Jensen et al.[30]. As a result, the overall correlation between predicted and observed retention times increased ($R=0.95$ and $R=0.98$; Fig. 3c, right) and the outlier peptides disappeared almost completely, leading to a lower $\Delta t_{95\%}$ value. While the refined retention time model was pre-trained and re-trained using tryptic peptides only, the prediction accuracy was also improved for non-tryptic peptides (Fig. 3d). Prosit always outperformed Elude[16], and predicted iRT values were closer to experimentally determined iRT values. This strongly suggests that Prosit learned general determinants of peptide retention and generalized across proteases. It also provides a means to generate a precise liquid-chromatography-specific iRT model for external data by transfer learning using a single LC–MS run provided by a user that represents the laboratory-specific liquid chromatography conditions, instead of requiring full re-training of Prosit on a comprehensive dataset (Supplementary Fig. 7).

Prosit spectral libraries can be used for DIA data analysis. The ability to predict iRT and fragment ion intensities for any peptide from sequence alone enabled the generation of spectral libraries for any organism in silico. To test their merits, we first obtained public spectral libraries for four different species (Orbitrap data) and predicted spectra for peptides contained in the respective libraries. As discussed above, the fragmentation prediction was calibrated to the external data, and the public spectral library was used for transfer learning iRTs (Supplementary Figs. 8 and 9a and Methods). The comparisons of the library spectra and Prosit's predictions resulted in almost identical spectral similarities (apex of spectral angle, ~0.9; $R>0.95$; Fig. 4a and Supplementary Table 4) for all four taxonomies.

Next, we benchmarked the in silico libraries against public datasets and their project-specific spectral libraries using Spectronaut. For evaluation of a HEK-293 dataset[31], the number of confidently identified peptide sequences resulting from exchanging the original library data with predicted iRT or intensity values was compared to the experimental library (filtered for long or modified peptides and fragment ions not considered by Prosit; Fig. 4b, Supplementary Fig. 9b,c and Supplementary Notes). Replacing iRT values led to the loss of 4,749 and a gain of 7,103 peptides. Exchanging fragment ion spectra showed a similar effect, and exchanging both retained 96.6% of the identifications of the original filtered library. The same was true on the protein level, where the number of identified protein groups using the filtered original library was 6,919, compared to 6,739 when using Prosit (Supplementary Fig. 9d). We repeated the above analysis for an Orbitrap DIA sample[31] containing proteins from *E. coli*, *S. cerevisiae* and *C. elegans* and obtained a very similar overall result compared to human proteins (Fig. 4b and Supplementary Fig. 9b–d).

To assess the transferability of Prosit predictions learned using Orbitrap data, we re-analyzed DIA-SWATH data of *S. cerevisiae*[10] and *D. melanogaster*[32] collected on quad time-of-flight (QTOF) instruments. The *S. cerevisiae* analysis followed the observed trend above, but part of the losses could be attributed to characteristics of TOF spectra (Supplementary Fig. 10 and Supplementary Notes). Using the predicted spectral library led to drastically more peptides in the *D. melanogaster* dataset (Fig. 4b). Close inspection showed that the predicted spectra far exceeded the quality of the spectra in the original library. The latter contained many low signal-to-noise spectra that did not accurately represent relative fragment ion intensities (Supplementary Fig. 10 and Supplementary Notes). These results underline the importance of generating high-quality and homogeneous spectral libraries for DIA analysis and suggest that Prosit provides a means to do this.

We integrated Prosit into ProteomicsDB[33,34] (https://www.proteomicsdb.org/prosit), allowing custom in silico spectral library generation (Fig. 4c,d). Users can upload FASTA files or a set of peptides and optionally calibrate NCEs and retention times against their own project-specific LC–MS/MS data to ensure high prediction quality. and ProteomicsDB will then perform and return predictions (Supplementary Fig. 11 and Supplementary Notes). In addition, we have integrated experimental and predicted reference spectra into ProteomicsDB, turning it into the largest and most comprehensive resource for high-quality reference spectra for any human peptide.

Intensity prediction improves the quality of peptide identification by database searching. A critical step in peptide identification by database searching is the removal of false (random) matches, which is typically achieved by controlling the false discovery rate (FDR) using the target-decoy approach[35]. We hypothesized that including fragment ion intensity information in the process of FDR estimation would increase the power of separating correct from incorrect identifications. To investigate this, we systematically compared acquired and predicted spectra from the Bekker-Jensen et al. tryptic dataset without applying any FDR filters (Methods). When we compared Andromeda scores with the corresponding spectral angle of the experimental and predicted spectra (Fig. 5a),

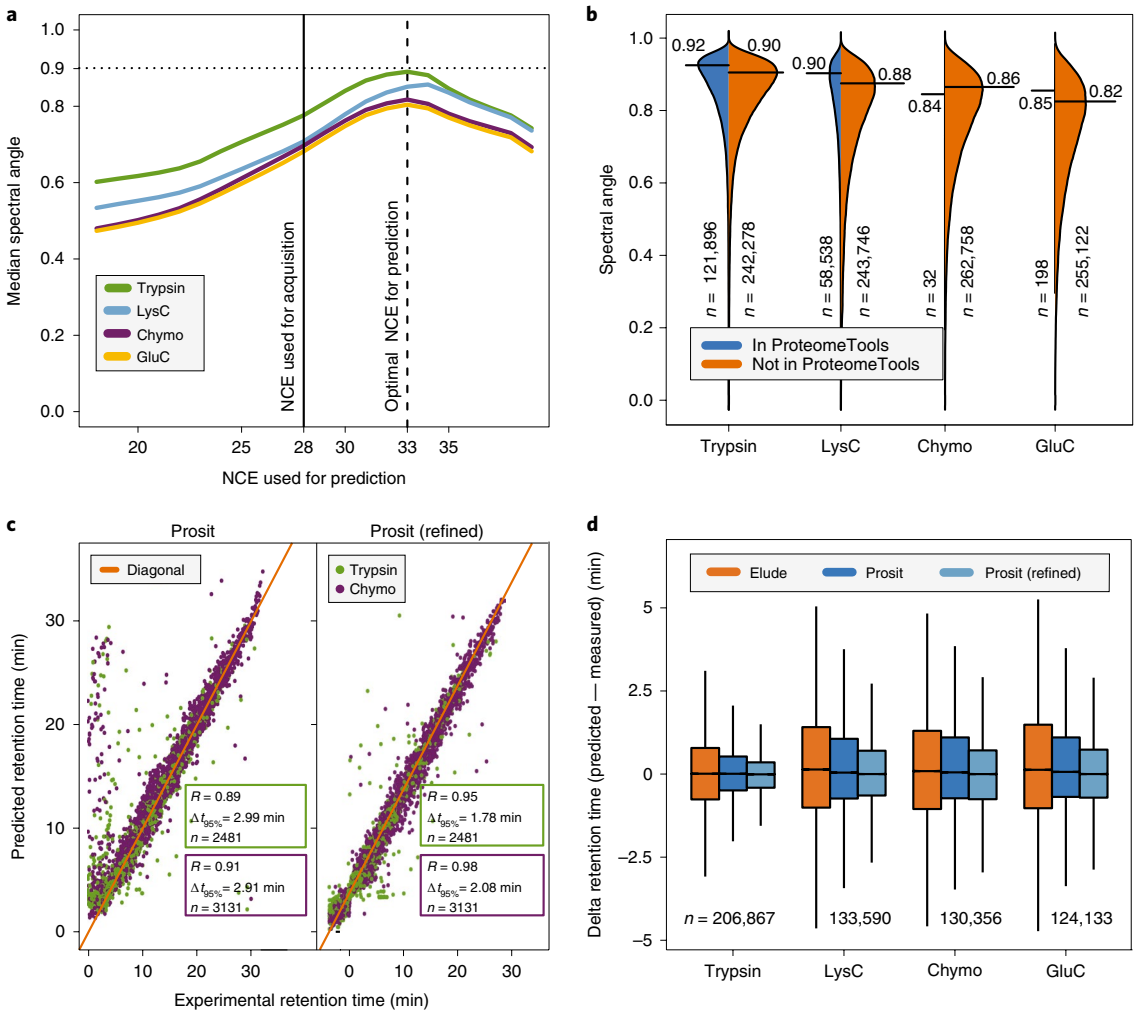

Fig. 3 | Evaluation of fragment ion intensity and iRT prediction for non-tryptic peptides. a, Median spectral angles resulting from a comparison of 1,000 high-scoring experimental spectra from the Bekker-Jensen et al. datasets to predictions at different NCEs for Trypsin (green), LysC (light blue), chymotrypsin (violet) and GluC (yellow). The solid black vertical line indicates the NCE used for acquisition, whereas the dashed vertical line indicates the NCE at which predictions best match the experimental spectra. **b**, Protease-specific spectral angle distributions of NCE-calibrated predictions for the Bekker-Jensen et al. dataset. The data were split between peptides also present in the ProteomeTools project (blue) or not (orange). Black lines and numbers indicate the apex of the spectral angle (SA) distributions. Distribution sample sizes are indicated. **c**, Examples of Pearson correlations (*R*) between experimentally determined and predicted retention times using either Prosit's general (left) or refined (right) iRT prediction model. Shown are two LC–MS/MS runs containing either tryptic (green) or chymotryptic (violet) peptides from the Bekker-Jensen et al. dataset ($n = 5{,}612$). The retention time window required to encompass 95% of all peptides ($\Delta t_{95\%}$) around the diagonal (orange line) is indicated by the blue lines. **d**, Box plots (no outliers shown) displaying the deviation between predicted and measured retention times for the Bekker-Jensen et al. dataset for each protease using Prosit's general (dark blue) and refined (light blue) iRT model, as well as for predictions performed by Elude (orange). The boxes indicate the IQR, and whiskers indicate 1.5× IQR values; no outliers are shown. The median delta retention time and number of peptides are indicated.

it was apparent that spectral angle separates target and decoy PSMs much more strongly than Andromeda scores. Of particular note, the exceptional accordance of the decoy distribution with the low-scoring target PSMs for the spectral angle suggests that the generation of decoy spectra is not biased, which is important for the correct estimation of false positive matches (Supplementary Notes). Although spectral angles and Andromeda scores correlated, the two scoring approaches frequently strongly disagreed (Fig. 5b). The peptide AQGLVTFR achieved an Andromeda score of 112, but the underlying spectrum showed very poor similarity to the predicted spectrum, as well as to the spectrum acquired for the synthetic peptide (spectral angle, 0.18; top spectra). In contrast, the spectra of the synthetic peptide and prediction agreed very well (spectral angle, 0.97; bottom spectra), suggesting that this peptide is a high-scoring false positive match. Conversely, the peptide LSGVEDHVK showed a rather low Andromeda score of 35 but a high spectral angle (spectral angle, 0.81). Here, both the spectrum of the synthetic peptide and prediction agreed well with the experimental spectrum (spectral angles of 0.81 and 0.96, respectively), suggesting that the peptide is a low-scoring but genuine identification, which would probably be removed at 1% FDR.

The above examples show that information about the presence/absence and abundance of fragment ions can provide additional information that may improve the scoring of database search results. To test this, we constructed a series of additional scores. For example, a simplified view on the Andromeda score is that it scores PSMs on the basis of the ratio of the number of observed versus all theoretically possible fragment ions of a peptide leading to a distribution of target and decoy identifications along the score axis (Fig. 5c, top). Doing the same but only using the fragment ions that Prosit predicts to be

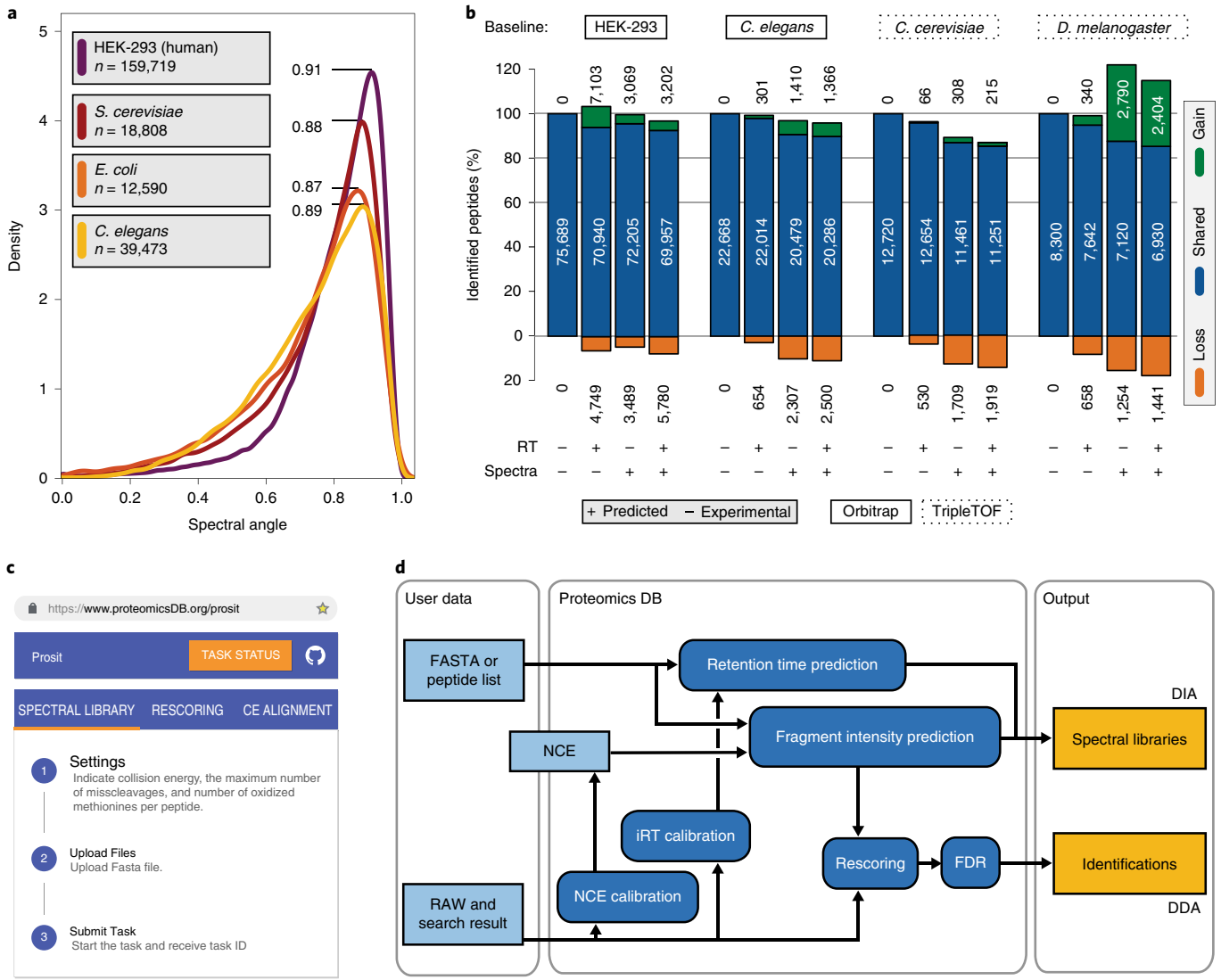

Fig. 4 | Prosit enables generation of in silico spectral libraries. a, Density distributions of spectral angles between predicted and experimental spectra from the HEK-293 (violet), *Saccharomyces cerevisiae* (red), *Escherichia coli* (orange) and *Caenorhabditis elegans* (yellow) Orbitrap spectral libraries. Numbers of spectral comparisons included in the distributions are indicated. **b**, Re-analysis of three data-independent acquisition (DIA)/SWATH datasets acquired using different mass analyzers using predicted spectral libraries (left to right: HEK-293, *C. elegans* and *Drosophila melanogaster*; Orbitrap, solid square; QTOF, dotted square). Data and project-specific spectral libraries were obtained from public repositories. For each organism, the original number of peptide identifications is shown on the left. The influence of gradually exchanging the original spectra or/and retention time information (as denoted by '+') by predicted values is shown in the following bars. The panel shows the number of retained peptides in blue, lost peptides in orange and gained peptides in green. **c**, Screenshot of the user interface of Prosit at http://www.proteomicsdb.org/prosit. **d**, Schematic representation of the integration of Prosit into ProteomicsDB for user access to predictions. Processing options include starting from a peptide list, a raw mass spectrometry file or search result files, to calibrate NCE and align iRT predictions of Prosit to their own data and to request the generation of predictions for a set of peptides, proteins or an entire FASTA file. Prosit can also rescore PSMs from data-dependent acquisition (DDA) experiments.

present in a spectrum (Fig. 5c, bottom) led to a much stronger separation between targets and decoys and a better estimation of the number of false matches in the low-scoring region. We have constructed several such new scores, capturing the peptide-, charge- and NCE-dependent number of observed versus predicted or observed but not predicted b and y ions (Supplementary Fig. 12, Supplementary Table 5 and Supplementary Notes). Subsequently, Percolator[36] was used to estimate the FDR for both the Andromeda score and the extended set of scores from Prosit to ensure a fair comparison (Methods).

Prosit scores greatly increased the separation between targets and decoy matches, as is apparent in the substantially increased number of targets at more stringent FDR cutoffs (Fig. 5d, top). The number of estimated true positive PSMs and peptides was remarkably constant across a wide range of FDR cutoffs, and the number of retained true positives at lower FDRs contained essentially all targets identified at 1% Andromeda score FDR (Fig. 5d, bottom). The improved scoring on the basis of Prosit extended to other proteases and increased the number of identified peptides between 5 and 35% at 1% peptide FDR (Fig. 5e, green bars). More surprisingly, using Prosit scores allowed the identification of the same or more peptides even at a 0.1% peptide level FDR compared to using Andromeda score at 1% peptide FDR (Fig. 5e, Supplementary Figs. 13 and 14 and Supplementary Notes), underscoring the substantial value of integrating further characteristics of tandem mass spectra into a scoring function. Performing a similar analysis using MS2PIP's predictions showed that Percolator was able to compensate for the

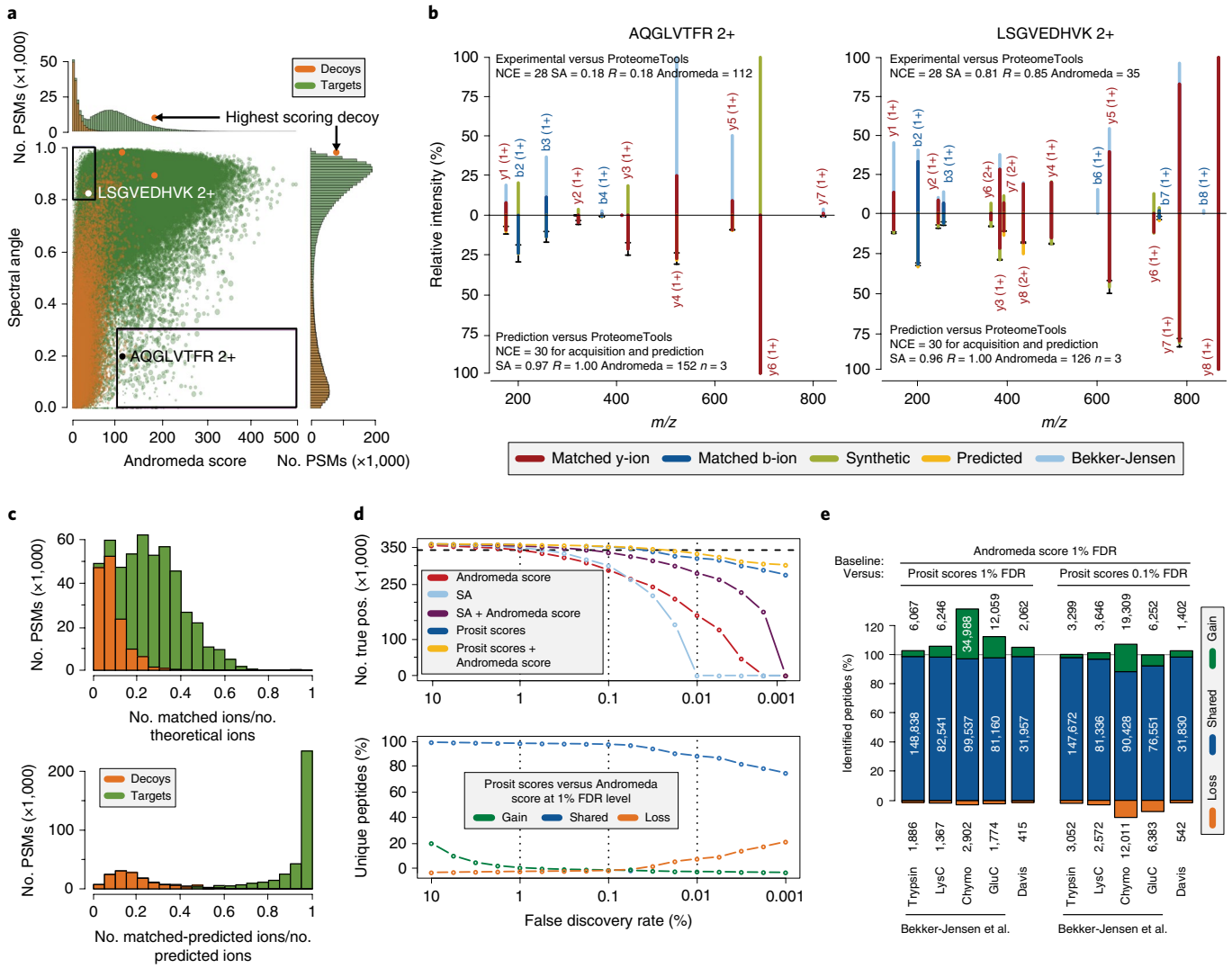

Fig. 5 | Intensity prediction greatly improves database search quality. a, Correlation of the Andromeda score to spectral angles for target (green) and decoy (orange) PSMs generated by Andromeda. Boxed regions indicate areas with strong disagreement as indicated by the low-scoring Andromeda but high-scoring spectral angle PSM for the peptide LSGVEDHVK (white) and high-scoring Andromeda but low-spectral-angle PSM for the peptide AQGLVTFR (black). The two PSMs highlighted in orange indicate the highest scoring decoy matches for Andromeda and spectral angle, respectively. The histograms on the axis show the separation of target and decoy PSMs. **b**, Spectra of the two sequences AQGLVTFR and LSGVEDHVK, also shown in **a**, are displayed as double mirror plots. The top panels compare the experimental spectrum from the Bekker-Jensen et al. dataset to the spectrum of the synthetic peptide from ProteomeTools at the NCE used for acquisition. The bottom panels compare the predicted spectra to the spectra of the synthetic peptide at the optimal NCE used for prediction estimated by NCE calibration. The left plot suggests a false positive identification of AQGLVTFR by Andromeda. The right plot suggests false negative identification of LSGVEDHVK by Andromeda. The coloring of the fragment ions visualizes the similarities (red, blue) and differences (green, orange, light blue) for each fragment ion. Black error bars indicate 1 s.d. around the measured fragment ion intensities and the color change between error bars the median. **c**, Examples of two peptide identification scores for target (green) and decoy (orange) PSMs. The one at the top is based on the number of all theoretically possible y- and b-type fragment ions. The number of matched non-zero fragment ions divided by the number of theoretical fragment ions approximates the Andromeda score. The score at the bottom is based on the number of the same fragment ions predicted to be present by Prosit. Here, the score uses the number of non-zero intensity fragment ions observed in the experimental spectrum and predicted to be present divided by the number of predicted non-zero ions. **d**, Comparison of the performance of a number of sets of scores used as Percolator input and evaluated by the number of true positive PSMs at different FDRs (top panel). The bottom panel shows the number of shared (blue), gained (green) and lost (orange) unique peptide identification using Prosit's set of scores at different FDR cutoffs compared to an Andromeda set of scores used as Percolator at 1% peptide level FDR. **e**, Peptide identification results using the Andromeda set of scores at 1% peptide level FDR are compared to results obtained with the Prosit set of scores for different external datasets. The left bar charts use a 1% FDR cutoff for both approaches. The right bar charts display the same analysis at 0.1% FDR cutoff for Prosit.

poorer prediction accuracy of MS2PIP and achieved a very similar increase in identified peptides on the tryptic dataset. However, for the chymotryptic dataset, MS2PIP reached only 70% of the identifications compared to Prosit's predictions (Supplementary Figs. 14 and 15 and Supplementary Notes).

Prosit improves the identification of peptides in metaproteomics data. The sensitivity of peptide identification by standard database searching is compromised when large sequence collections are investigated because higher and higher identification scores are required to satisfy a desired FDR[37]. This is particularly

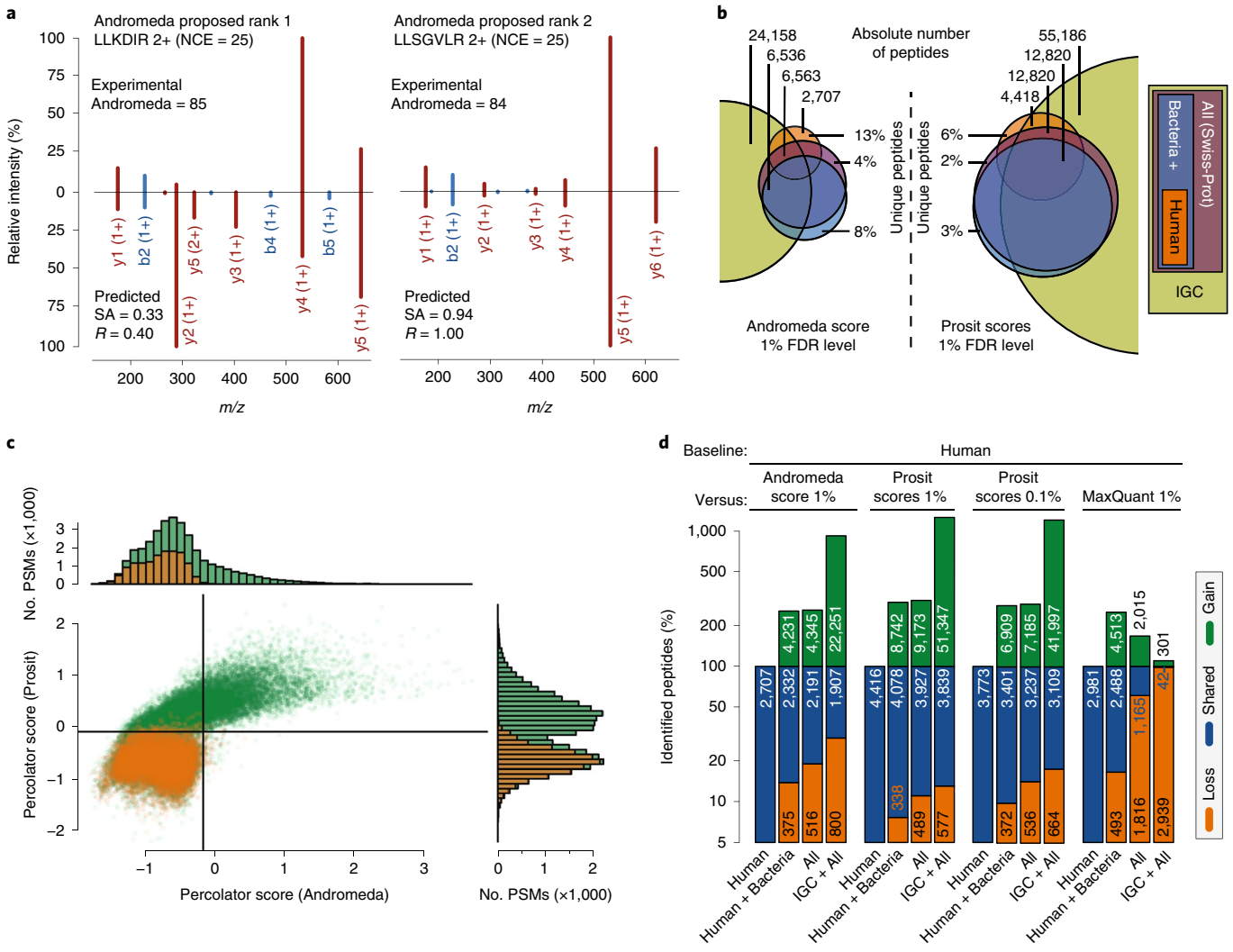

Fig. 6 | Prosit enables confident identification in large metaproteomic search spaces. a, Comparison of the predicted and experimental spectra of the top (left) and second (right) ranked peptides by Andromeda. The spectral comparison clearly shows that the second ranked peptide is the better match to the experimental spectrum. The color scheme is the same as in Fig. 2a. **b**, Results of database searching using metaproteomic samples against increasingly complex search spaces using Andromeda or Prosit at 1% peptide FDR. Absolute numbers in the Venn diagrams indicate the total number of peptides identified using the respective database. The relative numbers indicate the percentage of peptides exclusively identified when using the respective databases. **c**, Correlation of Percolator scores for all target (green) and decoy (orange) PSMs from using either the Andromeda or Prosit set of scores. The solid lines indicate the 1% peptide level FDR cutoffs for Andromeda and Prosit. **d**, Numbers of confidently identified peptides (*y* axis in $\log_{10}$) are shown for Andromeda, Prosit and MaxQuant at 1% peptide FDR (first, second and fourth panels) and Prosit at 0.1% peptide FDR (third panel). Sequence database sizes are increasing left to right in each panel (for details, see text). Retained, lost and gained peptide identifications compared to each search using human proteins only (baseline) are indicated by the blue, orange and green bars, respectively.

apparent in metaproteomics analysis where the search space is substantially enlarged by the many organisms that compose the sample[38]. To exemplify how Prosit improves peptide identification in this situation, Fig. 6a shows two PSMs matching to an experimental spectrum with very similar Andromeda scores (rank 1 top and rank 2 bottom). It is evident that Prosit's prediction for the second ranked peptide correlates more strongly with the experimental spectrum than the first (spectral angle, 0.94 versus 0.33; Methods), strongly suggesting that the second ranked peptide is indeed the correct identification.

To show that this is more generally true, we searched MS data from a human gut sample[39] against four increasingly complex sequence databases: (1) Swiss-Prot annotated proteins covering Human (20,260 proteins), (2) Human + Bacteria (276,628 proteins), (3) All organisms (469,313 proteins) and (4) IGC + All, which also incorporated the human gut microbial integrated gene catalog (IGC[40], 10,330,558 proteins and Fig. 6b). When we used Andromeda, up to 25% of identifications were lost as the search space became larger. The same analysis using Prosit scores led to the loss of only 13% of all identifications. As noted above, we again observed a much stronger separation of true and random matches (Fig. 6c), further demonstrating that the integration of intensity information led to a better use of the available information contained in tandem mass spectrometry spectra and hence more specific and comprehensive results. Only in rare cases did Andromeda identify peptides not identified by Prosit (Fig. 6c, bottom right corner), but Prosit confidently identified many tandem mass spectrometry spectra that did not pass Andromeda's scoring threshold (Fig. 6c, top left corner, and Supplementary Table 6) because Prosit was more confident in its top-ranking PSM (Supplementary Notes).

Using Prosit scores instead of Andromeda allowed Percolator to confidently identify up to 2.3 times more peptides than Andromeda (Fig. 6d). While both methods showed an increase in identifications when switching from Human to IGC + All, Andromeda lost 30% of the identifications compared to the Human database. These were not replaced by more confident sequences from bacterial peptides, but simply disappeared owing to their poorer *q* values resulting from increasingly overlapping target and decoy distributions. This was even more apparent at more stringent FDR cutoffs. At 0.01% FDR, Andromeda identified few if any peptides, whereas Prosit retained a large fraction of the identifications (Fig. 6d and Supplementary Notes). This shows that Prosit, at least partially, overcomes shortcomings of classical database searching. Here, this resulted in an overall increase of the identification rate to ~35% of all MS/MS spectra, which exceeds reported numbers for, as an example, two-step searches combining multiple search engines and smaller databases[41]. Even more striking in this context is that MaxQuant (without Percolator) on the same data led to the identification of only 343 peptides, compared to 55,186 peptides by Prosit (Fig. 6d).

Discussion

In this study, we introduce Prosit, a flexible deep neural network architecture able to predict retention times and tandem mass spectrometry spectra of peptides with near-synthetic peptide quality and that substantially surpass current benchmarks and tools. Although trained on tryptic peptides from human origin, it performed very well with all proteases, organisms, datasets, mass spectrometers and acquisition parameters tested here. This highlights that the learned internal representation of Prosit approximates a chemo-physical model for peptide fragmentation and chromatographic retention time. However, it is also clear that including more non-tryptic data or longer peptides as well as higher charge states would most probably further improve prediction accuracy.

Our results demonstrate that predicted spectral libraries can be used for analyzing DIA data. While predicted libraries performed slightly worse than high-quality experimental spectral libraries, replacing lower quality spectral libraries by consistent and high signal-to-noise predicted spectra increased the number of identified peptides by up to 20%. In the future, Prosit might enable the regeneration of libraries on instrument replacement or calibration and potentially supports the consistent addition of new peptide hypothesis without compromising the homogeneity of a library.

At this stage, Prosit requires a list of peptides expected in a sample from, for example, DDA experiments, because current DIA software cannot yet deal with spectral libraries containing entire (predicted) proteomes. However, we anticipate that these issues will be overcome in the near future (Supplementary Notes). Even though Prosit is not a full search engine, the incorporation of sequence-, charge- and NCE-dependent fragment intensities significantly improved classical database searching results because correct matches could be separated from incorrect matches more efficiently. This is of utmost importance when handling disproportionally large search spaces, as shown by the analysis of human feces[41]. This not only has major implications for metaproteomics, but will probably also improve other peptide-centric research areas such as proteogenomics[42] and immune peptidomics[43].

Although not demonstrated in this work, there is no technical or conceptual reason that Prosit's generic architecture could not be adapted to spectra acquired by other fragmentation techniques as used, for example, in the ProteomeTools project (that is, CID, electron-transfer dissociation (ETD), ETciD, EThcD), other fragment ion types (for example, neutral losses), modified peptides including post-translational modifications[44] and chemical labels. We envision that extension of Prosit in the future will lead to widespread adoption and use throughout bottom-up-based proteomics research. We are enabling this in a number of ways. First, all raw data acquired in the ProteomeTools project are available on PRIDE and spectral libraries are available on ProteomeTools.org and ProteomicsDB for all fragmentation techniques. Second, we have integrated predicted spectra at three practically useful NCEs for all human peptides into ProteomicsDB, allowing the validation of peptide identifications that are not covered by the ProteomeTools project. Third, the source code is available on Github and the current models of Prosit are available on figshare. Fourth, ProteomicsDB was extended to allow users to (1) calibrate Prosit predictions against custom data, (2) predict and download spectral libraries for any set of peptides or entire organisms and (3) run the peptide identification pipeline for single RAW files, covering all use cases presented in this manuscript.

We believe that the ability to predict fragment intensities with very high quality will make entirely new data processing options possible, supporting and enhancing proteomics research in the future. Beyond the examples shown above, Prosit can be envisaged to be useful for validating peptide identifications[45], multiple reaction monitoring (MRM)/parallel reaction monitoring (PRM) assay development, resolving chimeric spectra[46] and the integration of comprehensive proteome-wide predictions into bioinformatics workflows, to name a few. In any of these ways, we anticipate that Prosit will be of benefit for the scientific community and have an impact on the field of proteomics research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0426-7.

Received: 20 August 2018; Accepted: 18 April 2019;
Published online: 27 May 2019

References

1. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
2. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).
3. Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective. *Nat. Biotechnol.* **28**, 695 (2010).
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. Cox, J. et al. Andromeda: a peptide search engine integrated into the maxquant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
6. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
7. 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).
8. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
9. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
10. Schubert, O. T. et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat. Protoc.* **10**, 426–441 (2015).
11. Deutsch, E. W. et al. Expanding the use of spectral libraries in proteomics. *J. Proteome Res.* **17**, 4051–4060 (2018).
12. 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* **11**, O111.016717 (2012).
13. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
14. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **16**, 2246–2256 (2016).
15. Krokhin, O. V. & Spicer, V. Generation of accurate peptide retention data for targeted and data independent quantitative LC–MS analysis: chromatographic lessons in proteomics. *Proteomics* **16**, 2931–2936 (2016).

16. Moruz, L. et al. Chromatographic retention time prediction for posttranslationally modified peptides. *Proteomics* **12**, 1151–1159 (2012).
17. 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.* **22**, 214–219 (2004).
18. Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H. & Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput.* **2006**, 219–230 (2006).
19. Frank, A. M. Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* **8**, 2226–2240 (2009).
20. Degroeve, S., Maddelein, D. & Martens, L. MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Res.* **43**, W326–W330 (2015).
21. Zhou, X.-X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
22. Zolg, D. et al. PROCAL: a set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. *Proteomics* **17**, 1700263 (2017).
23. Zolg, D. P. et al. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262 (2017).
24. Wu, Y. et al. Google's neural machine translation system: bridging the gap between human and machine translation. Preprint at https://arxiv.org/abs/1609.08144 (2016).
25. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
26. Xu, K. et al. Show, attend and tell: neural image caption generation with visual attention. In *Proc. International Conference on Machine Learning* (eds. Bach, F. & Blei, D.) 2048–2057 (JMLR, 2015).
27. Krokhin, O. V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.* **78**, 7785–7795 (2006).
28. 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* **13**, 2056–2071 (2014).
29. Diedrich, J. K., Pinto, A. F. M. & Yates, J. R. Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot. *J. Am. Soc. Mass Spectrom.* **24**, 1690–1699 (2013).
30. Bekker-Jensen, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599 (2017).
31. Bruderer, R. et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics* **16**, 2296–2309 (2017).
32. Fabre, B. et al. Spectral libraries for SWATH-MS assays for *Drosophila melanogaster and Solanum lycopersicum*. *Proteomics* **17**, 1700216 (2017).
33. Schmidt, T. et al. ProteomicsDB. *Nucleic Acids Res.* **46**, D1271–D1281 (2017).
34. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
35. 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).
36. 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.* **27**, 1719–1727 (2016).
37. Shanmugam, A. K. & Nesvizhskii, A. I. Effective leveraging of targeted search spaces for improving peptide identification in tandem mass spectrometry based proteomics. *J. Proteome Res.* **14**, 5169–5178 (2015).
38. Muth, T., Benndorf, D., Reichl, U., Rapp, E. & Martens, L. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol. Biosyst.* **9**, 578–585 (2012).
39. Rechenberger, J. et al. Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. *Proteomes* **7**, 2 (2019).
40. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834 (2014).
41. Muth, T. R. et al. Navigating through metaproteomics data: a logbook of database searching. *Proteomics* **15**, 3439–3453 (2017).
42. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114 (2014).
43. Schumacher, F. R. et al. Building proteomic tool boxes to monitor MHC class I and class II peptides. *Proteomics* **17**, 1600061 (2017).
44. Zolg, D. et al. ProteomeTools: systematic characterization of 21 post-translational protein modifications by LC-MS/MS using synthetic peptides. *Mol. Cell. Proteomics* **17**, 1850–1863 (2018).
45. Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).
46. Dorfer, V., Maltsev, S., Winkler, S. & Mechtler, K. CharmeRT: boosting peptide identifications by chimeric spectra identification and retention time prediction. *J. Proteome Res.* **17**, 2581–2589 (2018).

Acknowledgements

This work was in part funded by the German Federal Ministry of Education and Research (BMBF, grant no. 031L0008A and no. 031L0168). The Titan Xp used in this research were donated by the NVIDIA corporation. The authors thank R. Bruderer (Biognosys) for sharing spectral libraries in textual and editable format, and R. Bruderer and members of the Kuster lab for fruitful discussions.

Author contributions

H.-C.E., S.A., B.K. and M.W. conceived the study. S.G., T.S., D.P.Z., J.R., K.S., J.Z., T.K., U.R., B.D., A.H., B.K. and M.W. designed experiments. S.G., T.S., D.P.Z., P.S., T.K., K.S., J.R., J.Z., B.D. and M.W. performed experiments. S.G., T.S., D.P.Z. and P.S. analyzed data. S.G., T.S., P.S. and M.W. extended the web resource. S.G., T.S., D.P.Z., B.K. and M.W. wrote the manuscript

Competing interests

M.W. and B.K. are founders and shareholders of OmicScouts. They have no operational role in the company. K.S., J.Z., T.K., H.W. and U.R. are employees of JPT. B.D. and A.H. are employees of Thermo Fisher Scientific. S.G., H.-C.E. and S.A. are employees of SAP SE.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0426-7.

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

Correspondence and requests for materials should be addressed to B.K. or M.W.

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

Methods

Synthetic peptides. *Generation.* To achieve extensive coverage of human proteins, four different sets of peptides were created and used in this study. First, we generated a large peptide set (termed 'Isoform') covering Swiss-Prot annotated isoforms by generating all feasible peptides (7–30 amino acids) distinguishing an isoform from the canonical form. The set comprises 123,514 peptides covering 9,354 isoforms. The second set, termed 'Missing gene addon', comprises 53,683 peptides (7–30 amino acids) and is extending the previously published 'Missing gene' set[40]. Peptide selection was analogous to that for the initial set and covers genes without sufficient evidence in ProteomicsDB[33,34]. Third, a set of 29,141 peptides covered sequences (7–40 amino acids) that are often detected in studies employing tandem mass tag labels. Note that the peptides are not modified with amine reactive isotope labels in this study. Fourth, we performed re-synthesis of 12,760 peptides that were part of the initial 'proteotypic' peptide set but were not detected using liquid chromatography–mass spectrometry (LC–MS). All sequences and protein mappings are available in Supplementary Table 1. Peptide Pool design, peptide synthesis, sample preparation and LC–MS of synthetic peptides are in detail described in ref. [23], including the Supplementary Information. In brief, peptide pools for synthesis and measurement contained roughly 1,000 peptides each. Near-isobaric peptides (±10 p.p.m.) were distributed across different pools of similar length to avoid ambiguous masses in pools wherever possible. All peptides were individually synthesized on cellulose membrane following the Fmoc-based solid phase synthesis strategy[47] using a purpose-built peptide synthesizer. The crude peptides were cleaved off the membrane in the predefined pools of 1,000 peptides and dried. Dried peptide pools were initially solubilized in 100% dimethyl sulfoxide (DMSO) to a concentration of 10 pmol µl^{-1} by vortexing for 30 min at room temperature. The pools were then diluted to 10% DMSO using 1% formic acid in high-performance liquid chromatography (HPLC)-grade water to a stock solution concentration of 1 pmol µl^{-1} and stored at −20 °C until use.

Data acquisition. Here, 10 µl of the stock solution were transferred to a 96-well plate and spiked with two retention time standards (Pierce Retention Time Standard and PROCAL) at 100 fmol per injection[22]. An estimated amount of 200 fmol of every peptide in a pool was subjected to liquid chromatography using a Dionex 3000 HPLC system (Thermo Fisher Scientific) using in-house-packed C18 columns. The setup consisted of a 75 µm × 2 cm trap column packed with 5-µm particles of Reprosil Pur ODS-3 (Dr. Maisch) and a 75 µm × 40 cm analytical column packed with 3-µm particles of C18 Reprosil Gold 120 (Dr. Maisch). Peptides were loaded onto the trap column using 0.1% formic acid in water. We separated the peptides by using a linear gradient from 4% to 35% acetonitrile with 5% DMSO, 0.1% formic acid in water over 50 min followed by a washing step (60 min total method length) at a flow rate of 300 nl min^{-1} and a column temperature of 50 °C. The HPLC system was coupled online to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). Each peptide pool was first measured using a 'survey method' consisting of an HCD (NCE, 28; Fourier transform mass spectrometry (FTMS)) and collision-induced dissociation (CID) (NCE 35, ion trap mass spectrometry (ITMS)) fragmentation event. From these identifications, three further methods were created to target only full-length synthesis products: (1) the '3×HCD' method comprised HCD events at NCE 25, 30, 35 (all FTMS); (2) the '2xIT_2xHCD' method comprised scans for CID NCE 35 ITMS, HCD NCE 28 ITMS, HCD NCE 20 FTMS, HCD NCE 23 FTMS; and (3) the 'ETD' method comprised an ETD and FTMS scan (charge depended reaction time), electron-transfer/collision-induced dissociation (EtciD) NCE 35 FTMS and electron-transfer/HCD (EthcD) NCE 28 FTMS. Aquired RAW data were analyzed using MaxQuant v.1.5.3.30 searching individual LC–MS runs against pool-specific databases (Supplementary Table 1). If not mentioned otherwise, default parameters were used: carbamidomethylated cysteine was specified as fixed modification, methionine oxidation as variable modification. The first search tolerance was set to 20 p.p.m., main search tolerance to 4.5 p.p.m. and filtered for PSM and protein FDR of 1%.

Synthetic peptide spectra are available on PRIDE with the dataset identifier PXD010595.

Statistics. *Spectrum similarity calculation.* We calculated spectrum similarity on the basis of all theoretically possible fragment ion intensities, while ignoring the *m/z* dimension. Let two spectra be S_a and S_b, with lengths n_a and n_b and precursor charges z_a and z_b represented by vectors V_a and V_b, respectively. V_a and V_b have the same length and contain the set of all y and b ion intensities in S_a and S_b up to ion $\max(n_a, n_b)-1$ for fragment charges up to $\min(\max(z_a, z_b),3)$ in the same dimension, respectively. Fragment ion intensities are base-peak normalized and intensities not observed or predicted are set to zero. For example, when S_a = PEPTIDE, $z_a = 2$ and S_b = PPTD, $z_b = 3$ then $n_a = 7$, $n_b = 4$ and V_a, V_b have length 18 ((7 length −1) × 3 precursor charge). We built this vector for all experimental, synthetic and predicted spectra and calculate similarity measures with it. For brevity in the following equations, let

$$\text{L2 norm}: \ |V|_2 = \sqrt{\sum_{i=0}^{n} |V_i|^2}; \ \text{mean deviation}: \tilde{V} = V - \frac{1}{n}\sum_{i=0}^{n} V_i; \ \text{L2 normed vector}: \hat{V} = \frac{V}{|V|_2}$$

We used the Pearson correlation (R) as implemented in scipy.stats.pearsonr (www.scipy.org) and defined as:

$$R(V_a, V_b) = \frac{\tilde{V}_a \cdot \tilde{V}_b}{|\tilde{V}_a|_2 \cdot |\tilde{V}_b|_2}$$

We defined normalized spectral angle (SA) as in ref. [28], but normalized vectors with an L2 norm as implemented in keras.backend.l2_normalize (www.keras.io). The loss function (normalized spectral contrast loss, SL) was defined as an adjusted spectral angle in the range 0 (high correlation) and 2 (low correlation), retaining its properties.

$$\text{SA} = 1 - 2\frac{\cos^{-1}(\hat{V}_a \cdot \hat{V}_b)}{\pi}; \ \text{SL} = 2\frac{\cos^{-1}(\hat{V}_a \cdot \hat{V}_b)}{\pi}$$

To avoid numerical instability during training on graphic processing units (GPUs) we added a fuzzing constant epsilon of 1×10^{-7} to all vectors.

Data distributions. Histograms shown depict data distributions. Statistical hypothesis tests, such as the *t*-test, were not used.

Fragmentation model. *Model architecture.* The peptide encoder consists of three layers: a bi-directional recurrent neural network (BDN) with gated recurrent memory units (GRU[48]), a recurrerrent GRU layer and an attention[49] layer all with dropout (Supplementary Fig. 2) The recurrent layers use 512 memory cells each. The latent space is 512-dimensional. Precursor charge and NCE encoder is a single dense layer with the same output size as the peptide encoder. The latent peptide vector is decorated with the precursor charge and NCE vector by element-wise multiplication. A one-layer length 29 BDN with GRUs, dropout and attention acts as decoder for fragment intensity (Supplementary Fig. 2). Implementation was done in Python with keras 2.1.1 and tensorflow 1.4.0 compiled to use GPUs. A keras model file can be found at https://github.com/kusterlab/prosit/.

Training data. For training, we used the publicly available ProteomeTools data (PRIDE datasets PXD004732 and PXD010595; see above). Data were searched using MaxQuant (v.1.5.3.30) with 1% FDR filter at PSM, protein or site level. The databases for the search are specific to the dataset, only containing the peptides to be expected in a specific pool, with carbamidomethylated cysteine specified as fixed modification and methionine oxidation as variable modification if not otherwise mentioned. Only the MaxQuant's top-ranking PSMs (msms.txt) were considered. MS2 spectra were extracted from the RAW files using Thermo Fisher's RawFileReader (http://planetorbitrap.com/rawfilereader). The extracted spectra were then annotated, whereas y and b ions were annotated at fragment charges 1 up to 3. Matching tolerances were 25 p.p.m. for FTMS and 0.35 Da for ITMS mass analyzers. We included all PSMs for the same peptide and restricted peptide length to 7–30 amino acids and precursor charge to <7 and Andromeda score to >50. Third, we transformed the annotation files to tensor format suitable for our machine learning models with a custom Python script. Ion intensities are continuous values and base-peak normalized. A spectrum is represented by a 174-dimensional vector (y/b ions, 3 charges, 29 fragment ions) and ordered as follows: y1 (1+), y1 (2+), y1 (+3), b1 (1+), b1 (2+), b1 (2+), y2 (1+) and so on. Fragment ion intensity values at impossible dimensions (that is, y20 for a 7-mer) are set to −1. All NCE values used during acquisition were aligned to a reference dataset to allow consistent NCE prediction, as described in ref. [22]. Briefly, for PROCAL RT peptides, HCD spectra at one NCE acquired in the ProteomeTools project were compared to spectra acquired at 15 NCEs from ref. [22]. A calibration curve was established whose intercept was used to offset the NCE in the ProteomeTools project. The offset value was then used as input to Prosit.

Inputs to the fragmentation models are peptide sequence, precursor charge and NCE. Peptide sequences are represented as discrete integer vectors of length 30, with each non-zero integer mapping to one amino acid and padded with zeros for sequences shorter than 30 amino acids. Precursor charge is one-hot encoded and NCE is normalized to [0, 1].

Training. Training data were split into three distinct sets with each peptide sequence included in only one of the three: 'Training' (72%, ~331,000 peptides, ~5.43 million PSMs), 'Test' (18%, ~82,000 peptides, ~1.36 million PSMs) and 'Holdout' (10%, ~46,000 peptides, ~0.75 million PSMs). For training, the data were restricted to a maximum of three PSMs per precursor (peptide sequence, modifications, charge) with an Andromeda score of >100, and decoy hits were excluded. The model was trained and optimized on Training. Test was used to control for overfitting with early stopping. The Holdout dataset was used to evaluate the model's generalization and potential biases. Normalized spectral contrast loss (see Spectrum Similarity Calculation) was used as a loss function. We used the Adam optimizer with an initial learning rate of 0.001 and 512 samples per batch. We trained on Nvidia TitanXp GPUs for 32 epochs. During training, fragment ions out-of-range or out-of-charge are masked and not regarded in further analysis. For example, for a length 10 peptide y and b ions 10–29 are masked and fragment ions with a charge higher than its precursor. PSMs with two or less matched fragment ions were discarded.

We provide all models and training data at https://figshare.com/projects/prosit/35582.

Calibration. To align Prosit to experimental data, up to 10,000 PSMs were randomly sampled from a dataset and for each peptide sequence spectra for all NCEs between 20 and 40 were predicted. Subsequently, all experimental spectra were compared to each of the predicted and the calibrated NCE was the NCE with the highest median spectral angle. We used PSMs from the PROCAL paper[22] to validate inter- and extrapolation performance of Prosit. It includes 40 peptides measured at 15 different NCEs. Prosit's training, test and holdout datasets cover only five different NCEs and do not include the 40 PROCAL peptides.

Retention time model. *Model architecture.* The peptide encoder is identical to the peptide fragmentation model. The latent space is a 512-dimensional vector that is fully connected to a one-dimensional output layer. Implementation was done with the same Python, keras and tensorflow implementation as the fragmentation prediction model and is also available at https://github.com/kusterlab/prosit/.

Training. For training, the publicly available ProteomeTools data (PRIDE dataset PXD004732 and PXD010595; see above) were split into three distinct sets similar to the fragmentation prediction training. The two most stable retention time spike-in peptides were chosen as retention time pillars to project all retention times in a unitless indexed retention time (iRT) space. The reference peptide eluting earlier was assigned an iRT value of 0 ('ISLGEHEGGGK') and the other one an iRT value of 100 ('GFVIDDGLITK'). iRT values for all other peptides were then calculated on the basis of a linear regression between the two pillars. Data for learning was restricted to peptides identified with a minimum MaxQuant score of 70 and an iRT variance between different ProteomeTools runs of less than 1. The remaining data were split into 64% training data, 16% test data and 20% holdout data. The mean squared error was used as loss function. The same data representation, infrastructure and optimizer was used as for training the fragmentation model.

Prediction. Because Prosit provides iRT values for peptide sequences and not experimental retention times, we used a customized loess fit to align our iRT space to experimental retention time. For this purpose, we used a non-linear regression calibration. We use a loess fit implementation in the R package fAncova to control for an optimal smoothing parameter. The optimal loess fit is found if the median of the absolute prediction error is smaller than 0.05. If the error is bigger, the one per million quantile of the absolute prediction error is calculated and all values are discarded showing a higher error. This process is applied recursively until the stop criterion is reached. The implementation of the loess fit can be found at https://github.com/kusterlab/prosit/. For RAW-file specific calibration and evaluation, we used high-confidence peptide identifications (that is, $q < 0.01$ for DDA). In case multiple elution profiles were detected for one peptide, the highest scoring (that is, Andromeda score) was used as the representative. Calibration and evaluation for spectral libraries was performed on all provided iRTs.

Model refinement by transfer learning. For transfer learning, either experimental retention times (tryptic DDA only with an Andromeda score > 50) or iRTs provided in the spectral libraries (HEK-293, *E. coli*) were used. Subsequently, the (indexed) retention times were scaled to be centered at 0 with a standard deviation of 1 (*z*-scoring). Then 80% of this dataset was used to refine the existing Prosit model and the remaining 20% was used to control for overfitting. Prosit was initialized with the model weights trained on the ProteomeTools data. The refined iRT model was then used to predict iRT values as described earlier.

Elude. The highest scoring peptide identifications were used to train RAW-file specific Elude (v.3.02) (refs. [16,50]) models using the standard settings. If necessary, the training data for Elude were sub-sampled to 1,500 PSMs to control for training time.

We provide all models and training data at https://figshare.com/projects/prosit/35582.

Processing of external data. *Data processing.* The Bekker-Jensen et al.[30] multi-protease dataset was downloaded from the PRIDE repository with the identifier PXD004452. The Davis et al.[51] dataset was downloaded from the PRIDE repository with the identifier PXD003977. We searched the RAW files using MaxQuant (v.1.5.3.30) with FDR filter 100% at PSM, protein or site level to generate complete candidate peptide lists for all scans in each RAW file. A human Swiss-Prot protein sequence database including annotated isoforms (downloaded 2 July 2016; 42,164 protein sequences) was used for MaxQuant processing; all further settings were left as preconfigured. Only the MaxQuant's top-ranking PSMs (msms.txt) were considered. MS2 spectra were extracted from the RAW files using Thermo Fisher's RawFileReader (http://planetorbitrap.com/rawfilereader). The y and b ions of the extracted spectra were annotated at fragment charges 1 up to 3. Matching tolerances were 25 p.p.m. for FTMS and 0.35 Da for ITMS mass analyzers. We included all PSMs for the same peptide and restricted peptides length to 7–30 amino acids and precursor charge to <7. Annotation files were transformed to tensor format suitable for our machine learning models with a custom Python script. Ion intensities are continuous values and base-peak normalized. Spectra comparison was performed on annotated y and b ions only.

FDR calculation. We used Percolator 3.01 (https://percolator.ms) with its standard settings for FDR calculation, but explicitly specifying target-decoy-competition (-Y flag) and retaining redundant peptides (-U flag). For datasets with more than one RAW file, we constructed a scan number column that is RAW-file specific. We compared different Percolator input files on the basis of different feature sets for all datasets, each including all default features recommended. The different score sets are described in detail in Supplementary Table 5. Briefly, the Andromeda score set extends the default features of Percolator by Andromeda score and Andromeda delta score. The *SA* set extends the default features by spectral angle. The *SA*+ Andromeda score set combines the former two score sets. Prosit scores extend the default features by a variety of scores on the basis of predicted fragment ion intensities (see Supplementary Table 5). Note that whenever we mention a score set name in the context of FDR analysis, Percolator was used with that score set to calculate FDR. Specifically, mentioning 'Andromeda score' in Figs. 5d,e and 6d means using Percolator with the Andromeda score set. Mentioning 'Prosit scores' in those figures means using Percolator with Prosit scores. 'MaxQuant 1%' in Fig. 6e is an exception: it reports the FDR calculated by MaxQuant for a comparison to the Percolator-based 'Andromeda 1%'. All results generated in this context are available on PRIDE with the identifier PXD010871.

DIA data analysis. For the reprocessing of public data-independent acquisition (DIA) datasets, multiple datasets and belonging spectral libraries were downloaded. For the HEK-293/HeLa/multi-organism study by Bruderer et al.[31], the RAW files were obtained from PRIDE (PRIDE identifier PXD005573). The spectral library in textual format was shared by the authors of the study. To filter the library for peptide entries not yet supported by Prosit, we removed all modified peptides (besides methionine oxidation), neutral loss fragment ions and kept only peptides between 7 and 30 amino acids in length with a precursor charge of 1–6. We used the method described in NCE calibration on a DDA file from the same study to calibrate fragment intensity predictions. An aligned collision energy of NCE 33 matched best to the acquired spectra in the library. To align the predicted indexed retention time, we used loess fit to align the predicted iRT to the iRT values present in the spectral library. We further tuned (transfer learning) the iRT prediction of Prosit by refining the model with the presented iRT values. Therefore, all iRT values were scaled to a mean of 0 and a standard deviation of 1 and 80% of the data was then used for refinement of Prosit while 20% was used to control for overfitting.

For the generation of spectral similarity distributions between predicted spectra and obtained spectral libraries (Fig. 4a and Supplementary Fig. 10a), predicted spectra were compared to their experimental counterparts from the respective spectral library using spectral angle as described above. All non-zero fragment ions ($m/z > 300$, ion >3, no neutral loss fragment ions) were considered for spectral angle calculation. For the generation of spectral similarity distributions (Supplementary Fig. 10d) between predicted spectra and *S. cerevisiae* DDA QTOF spectra, the .wiff files were searched using MaxQuant v.1.5.3.30 using standard settings for TOF spectra as well as 1% peptide and 1% protein FDR against the *S. cerevisiae* UniProt reference proteome. Deisotoping of TOF spectra was deactivated. Predicted spectra were compared against the annotated fragment ions from the MaxQuant msms.txt (b and y ions considered, no neutral loss fragments) using SA.

For DIA analysis (Fig. 4b, Supplementary Figs. 9 and 10 and Supplementary Notes), the spectral libraries were imported into Spectronaut 11 (v.11.0.15038) using slightly modified standard settings: m/z minimum 300, m/z maximum 1,800, minimum relative intensity 5%, best six (minimum 6) fragment ions per spectrum, ion >3. Respective protein databases for spectral library import and DIA processing were the following: Human Swiss-Prot including isoforms (downloaded 20 July 2016; 42,164 protein sequences), *C. elegans* UniProt reference proteome UP000001940 (downloaded 11 July 2018; 26,796 protein sequences), *S. cerevisiae* UniProt reference proteome UP000002311 (downloaded 11 July 2018; 6,048 protein sequences) and *E. coli* K12 UniProt reference proteome UP000000625 (downloaded 11 July 2018; 4,313 protein sequences). Spectronaut processing was performed with slightly modified standard settings: automatic iRT calibration, XIC RT width automatic, MS1 and MS2 mass tolerance dynamic, decoy limit dynamic, minimum 10% or 5,000 entries, decoy generation method 'mutated', precursor and protein q value cutoff at 0.01. Standard reports for all searches were generated and the unique peptide and protein entries per condition per replicate counted and compared. Visualization was performed using custom R scripts. The bars displayed in Fig. 4b and Supplementary Fig. 10 are median aggregate values over three replicates. The Loss and Gain parts of the bars sum to the same value in the same group of bars. For the aggregations over multiple replicates displayed, this may not be the case and can lead to small divergences.

For the *S. cerevisiae* QTOF study (TripleTOF 5600) by Schubert et al.[10], the .wiff files and spectra library were obtained from the PRIDE repository (PRIDE identifier PXD001126). The DDA RAW files were searched using MaxQuant v.1.5.3.30 using standard settings for TOF spectra as well as 1% peptide and 1% protein FDR against the *S. cerevisiae* UniProt reference proteome. The spectra

library was created by Spectronaut using the settings above. Spectra for precursors contained in the library were predicted using an NCE of 29, and we further tuned (transfer learning) the iRT prediction of Prosit by refining the model with the presented iRT values. Spectronaut processing was with the settings stated above, using the *S. cerevisiae* UniProt reference proteome UP000002311 (downloaded 11 July 2018; 6,048 protein sequences).

For the *D. melanogaster* study (TripleTOF 6600) by Fabre et al.[32], the RAW files and spectra library were obtained from the MassIVE repository (MassIVE identifier MSV000081075). Spectra were predicted using an NCE of 30, iRT was aligned as stated above using a custom loess fit and no further refinement of the iRT model was applied. Spectronaut processing was with the settings stated above, using the *D. melanogaster* UniProt reference proteome UP000000803 (downloaded 8 May 2018; 23,297 protein sequences).

Spectronaut search files, search reports and spectral libraries are available on PRIDE with the identifier PXD010871.

Metaproteomics. *Data acquisition.* LC–MS/MS measurement was performed using a Dionex Ultimate 3000 UHPLC+ system coupled to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific). After reconstitution in 0.1% formic acid, peptides were delivered to a trap column (75 µm × 45 cm, packed in-house with 5-µm C18 resin; Reprosil PUR AQ, Dr. Maisch) and washed using 0.1% formic acid at a flow rate of 5 µl min^{-1} for 10 min. Subsequently, peptides were transferred to an analytical column (75 µm × 2 cm, packed in-house with 3-µm C18 resin; Reprosil PUR AQ, Dr. Maisch) with a flow rate of 300 nl min^{-1} and separated using a 60-min gradient from 4% to 32% liquid chromatography solvent B (0.1% formic acid, 5% DMSO in acetonitrile) in liquid chromatography solvent A (0.1% formic acid, 5% DMSO). The instrument was operated in data-dependent acquisition (DDA) and positive ionization mode. MS1 full scans were acquired from 360 to 1,300 *m/z* at a resolution of 60 K, an automatic gain control (AGC) target value of 3×10^6 charges and a maximum injection time (maxIT) of 10 ms. Precursor ions for HCD fragmentation were selected with a top 20 method and MS2 spectra were recorded from 200 to 2,000 *m/z* at 30 K resolution using an isolation window of 1.7 *m/z*, an AGC target value of 2×10^5 charges, a maxIT of 50 ms, NCE 25, a dynamic exclusion of 20 ms and a fixed first mass of 100 *m/z*.

Data preprocessing. Data were initially searched using MaxQuant (v.1.5.7.4) with FDR filter 100% at PSM, protein or site level to generate a complete candidate peptide list for all scans in each RAW file. The four used databases were Swiss-Prot human (downloaded 20 July 2016; 20,260 protein sequences), Swiss-Prot human + all bacteria (concatenation of human + 276,628 protein sequences), Swiss-Prot all organisms (downloaded 20 July 2016; 469,313 protein sequences) and IGC[40] (concatenated all + 10,330,558 protein sequences). Additional parameters were as described earlier. We used the same data extraction as described in external data preprocessing, but instead of including only the top-ranking PSM, we included up to 15 PSMs identified by MaxQuant per scan. For this, we extracted all PSMs per scan from the Andromeda generated res-files (andromeda folder). Prediction, comparison and FDR calculation was performed as described in external data processing. Besides FDR calculation, Percolator was used to also select the top-ranking PSM per scan on the basis of its meta-score optimization. A second processing of the RAW files was performed using MaxQuant's partial processing option, setting the PSM FDR level to 1%. Note that 'MaxQuant 1%' in the Fig. 6d FDR calculation comes directly from MaxQuant, in contrast to 'Andromeda 1%' in the same figure, which comes from Percolator with the Andromeda score set.

All results generated in this context are available on PRIDE with the identifier PXD010871.

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

Data availability

Reference spectra are available at https://www.proteomicsdb.org, and updates to the resource are available at http://www.proteometools.org. The mass spectrometric raw data of ProteomeTools have been deposited with the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD010595. The MaxQuant and Spectronaut search data including intermediate results underlying the presented analysis have been deposited with the dataset identifier PXD010871. Learned Prosit and Elude models are deposited at https://figshare.com/projects/prosit/35582.

Code availability

Source code and scripts are available on GitHub at https://github.com/kusterlab/prosit.

References

47. Wenschuh, H. et al. Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Pept. Sci.* **55**, 188–206 (2000).
48. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. Preprint at https://arxiv.org/abs/1412.3555 (2014).
49. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. Preprint at https://arxiv.org/abs/1409.0473 (2014).
50. Moruz, L., Tomazela, D. & Käll, L. Training, selection, and robust calibration of retention time models for targeted proteomics. *J. Proteome Res.* **9**, 5209–5216 (2010).
51. Davis, S. et al. Expanding proteome coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) combined with broad specificity proteolysis. *J. Proteome Res.* **16**, 1288–1299 (2017).

nature research

Corresponding author(s): *Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	After acquisition, data was processed with MaxQuant (1.5.3.30) and stored in a custom PostgreSQL (9.5.7) database. Data extraction was performed using custom bash (4.3.48), R (3.4.0) and SQL (9.5.7) scripts.
Data analysis	The machine learning model was implemented in Python with keras (2.1.1), tensorflow (1.4.0), numpy (1.14.5) and scipy (1.1.0) and compiled to use graphic processing units (GPU). Source code is available at www.github.com/kusterlab/prosit/. Evaluated models in this study are available at figshare.com/projects/prosit/35582. Further software used: Thermo RawFileReader (3.0.54), Elude (3.02), SSRCalc (Q.0), MS2PIP (Server: https://iomics.ugent.be/ms2pip/ accessed Mar-Nov 2018), Percolator (3.01), R (3.4.0) package fAncova, Spectronaut 11 (11.0.15038), pDeep (http://pfind.ict.ac.cn/download/pDeep.zip accessed Mar 2018)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Mass spectrometric RAW data is available on ProteomeXchange with the IDs PXD010595 and PXD010871 . Public datasets used for re-processing are PXD004732 (training data), PXD004452 (Figure 2, 3, 5), PXD005573 (Figure 4), MassIVE repository identifier MSV000081075 (Figure 4), PXD003977 (Figure 5).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Available training data was split into three distinct sets with each peptide sequence only included in one of the three: “Training” (72%, ~331,000 peptides, ~5,430,000 PSMs), “Test” (18%, ~82,000 peptides, ~1,360,000 PSMs) and “Holdout” (10%, ~46,000 peptides, ~750,000 PSMs).
Data exclusions	For training, the data was restricted to maximum 3 PSMs per precursor (peptide sequence, modifications, charge) with an Andromeda score >100 and decoy hits were excluded to 1) allow training data to fit into main memory of the computer and 2) filter for high quality reproducible spectra.
Replication	Training was performed 5 times with a random split of the available training data each time.
Randomization	"Training", "Test" and "Holdout" were generated using random splits of the available training data, with the contraint that a peptide sequence was only allowed to be present in either of the three groups.
Blinding	Blinding is not a relevant technique in machine learning. To avoid over-fitting we applied three regularization techniques: early stopping, dropout and the use of noisy data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging