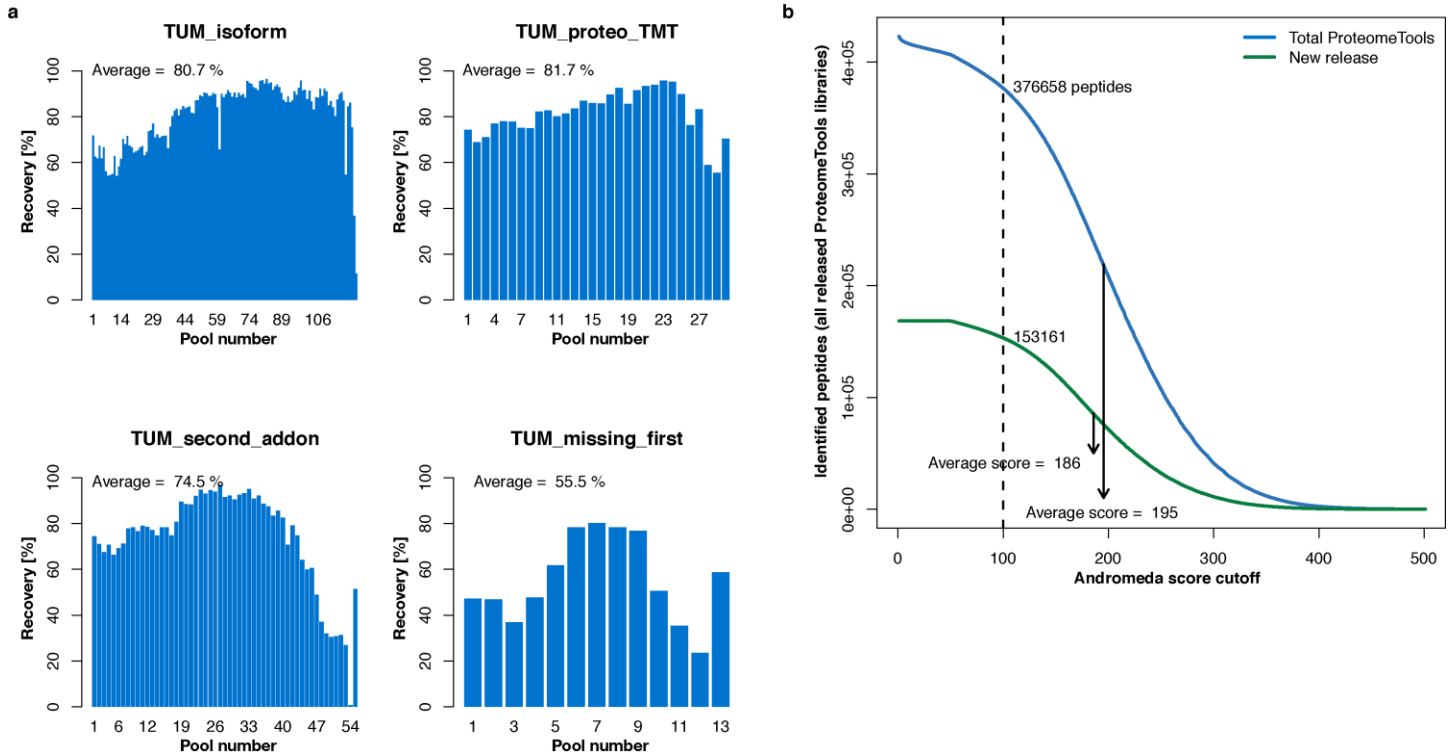


In the format provided by the authors and unedited.

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

Siegfried Gessulat ^{1,2,7}, Tobias Schmidt^{1,7}, Daniel Paul Zolg¹, Patroklos Samaras ¹,
Karsten Schnatbaum³, Johannes Zerweck³, Tobias Knaute³, Julia Rechenberger¹, Bernard Delanghe⁴,
Andreas Huhmer⁵, Ulf Reimer³, Hans-Christian Ehrlich², Stephan Aiche ², Bernhard Kuster ^{1,6*}
and Mathias Wilhelm ^{1*}

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

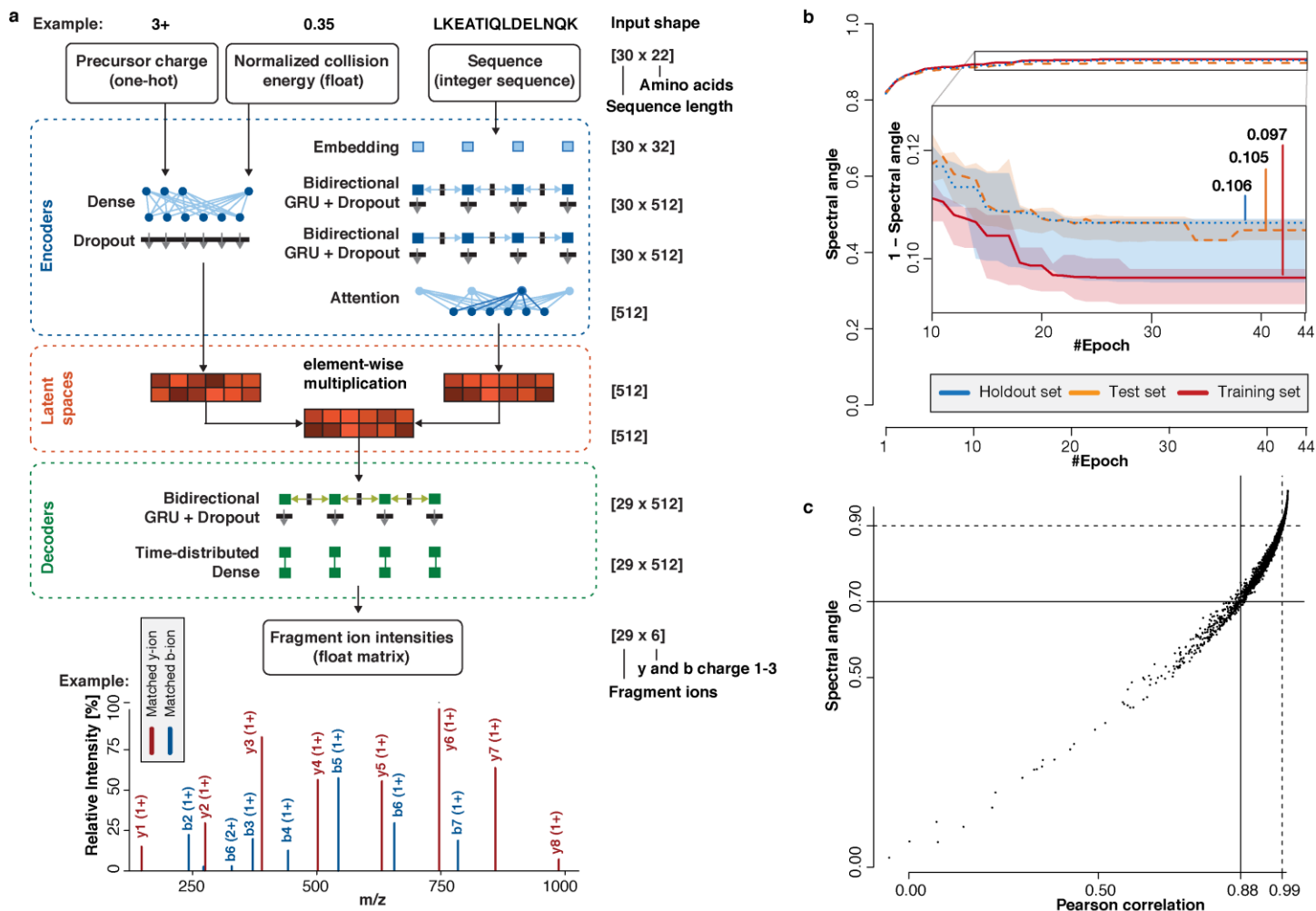


Supplementary Figure 1

Overview of identified peptides in the ProteomeTools project

(a) Recovery of synthesized peptide sequences across all four new datasets. Bars display the percentage of peptides identified in comparison to the peptides synthesized per pool of ~1000 peptides. Only identifications with an Andromeda Score of at least 50 are considered.

(b) Identified peptides over Andromeda score cutoff for both the newly released dataset as well as the complete ProteomeTools peptide library. Numbers at the arbitrary cutoff of 100 are displayed for both datasets, the median Andromeda Score is indicated.



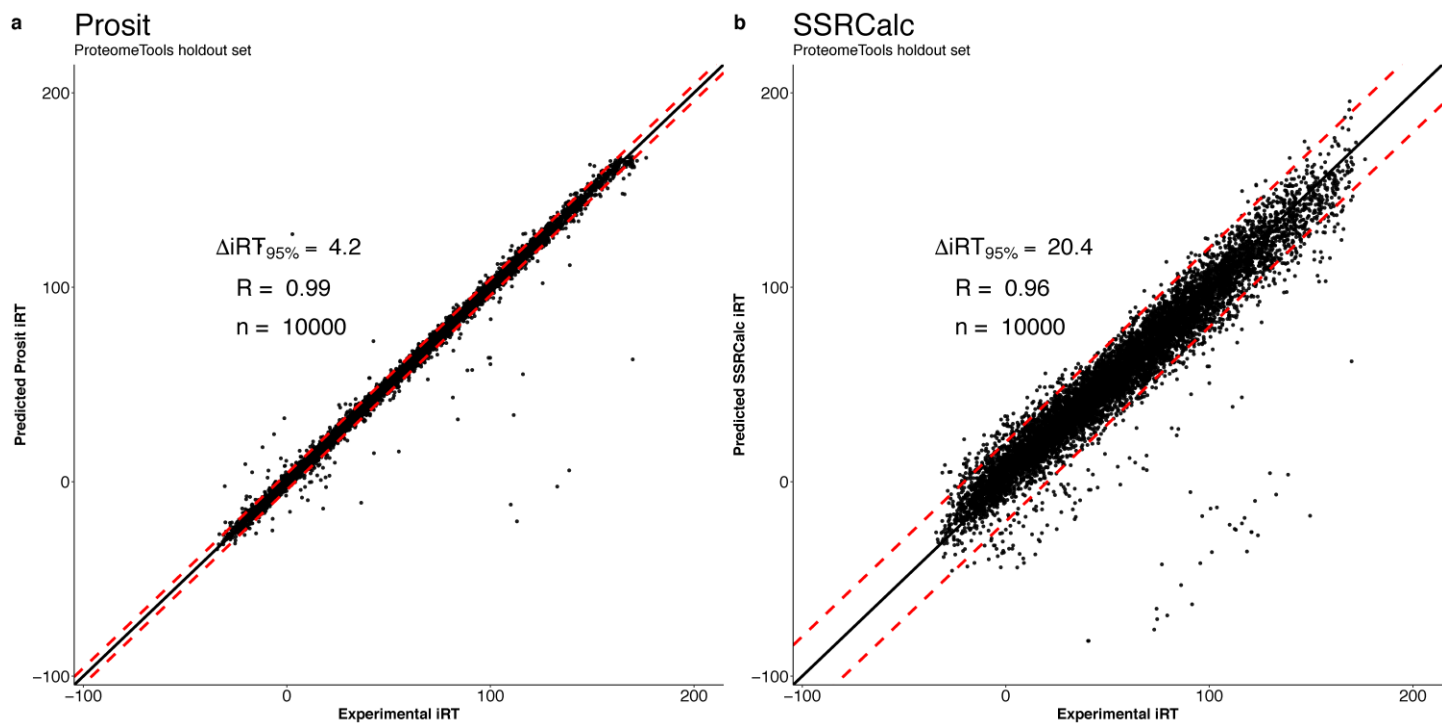
Supplementary Figure 2

The Prosit deep learning model and its training

(a) Overview of the neural network architecture of the fragment ion intensity prediction model. The model takes precursor charge, normalized collision energy and the peptide sequence as input. First, for every input a specific encoder is trained, consisting of one dense layer for precursor charge and normalized collision energy. The encoder for the peptide sequence is split in an embedding layer connected to 2 bi-directional recurrent neural networks (BDN) with gated recurrent memory (GRU) units and an attention layer. Both encoder representations are element-wise multiplied for a fixed size latent space representation. The decoder for fragment ion intensity prediction consists of one bidirectional GRU resulting in 6 predictions for up to 29 fragmentation positions. The indexed retention time (IRT) model uses the same encoder but dense layers as decoder.

(b) Model performance for 5 random splits of the ProteomeTools data into Training, Test and Holdout. The main panel shows best performing models from 5 random splits of the data. The inset details the median models error with intervals (shaded regions) ranging from the best performing model to the worst performing model over the 5 splits for Training Test and Holdout.

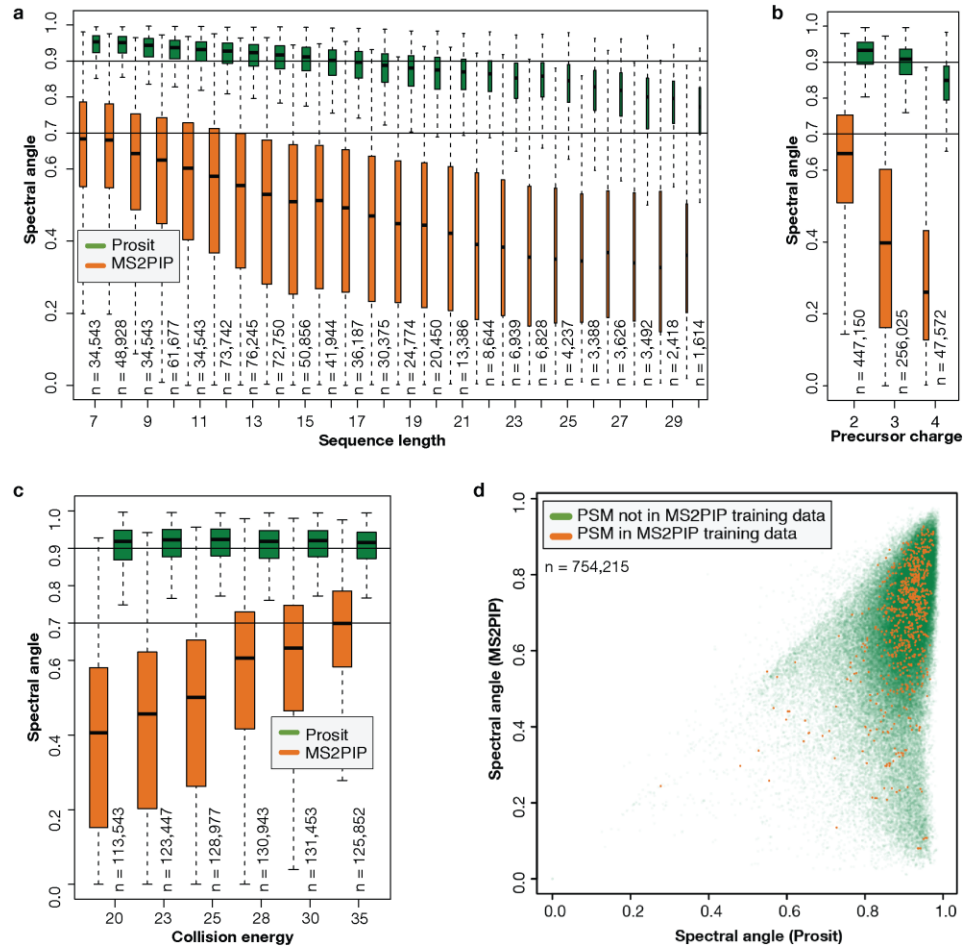
(c) Comparison of Pearson correlation and normalized spectral contrast angle (short spectral angle) as measures for spectral similarity between predicted and measured spectra contained in the holdout set for fragment ion intensity prediction.



Supplementary Figure 3

iRT prediction using SSRCalc and Prosit on the ProteomeTools holdout set

Benchmark of the indexed retention time (iRT) prediction model of Prosit **(a)** in comparison to SSRCalc **(b)**. Plotted are the predicted and measured iRT values of peptides (dots) in the holdout set. The required iRT window that would encompass 95% of all peptides is indicated by the red dashed lines.



Supplementary Figure 4

Comparison of Prosit and MS2PIP

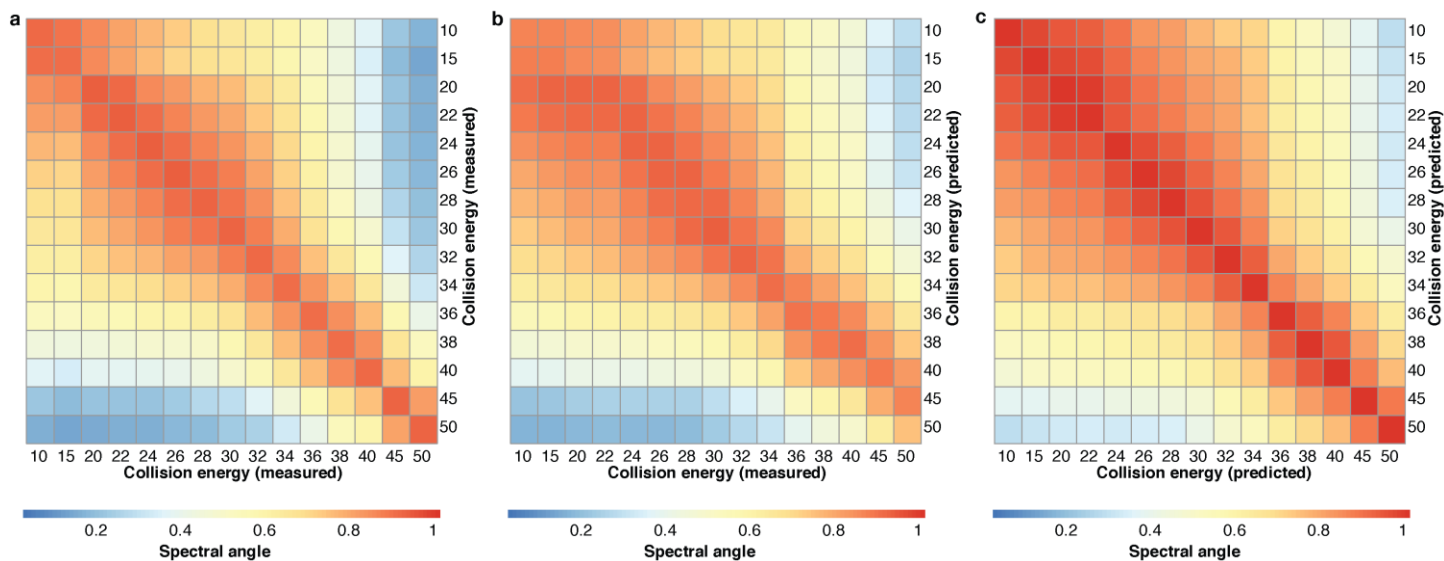
Subsets of the ProteomeTools holdout are used in this figure. None of the peptides and spectra in this dataset were used to train or test Prosit's fragment intensity model. In boxplots, outliers are not shown, whiskers indicate 1.5 interquartile range (IQR), and black horizontal lines indicate median values. For reference, a spectral angle of 0.9 and 0.7 are indicated.

(a) Benchmark of Prosit's (green) and MS2PIP's (orange) fragment ion intensity prediction compared to the experimental ProteomeTools spectrum respectively. Data is split by peptide length on a random subset of the ProteomeTools holdout dataset.

(b) Same as (a) but split by precursor charge

(c) Same as (a) but split by collision energy

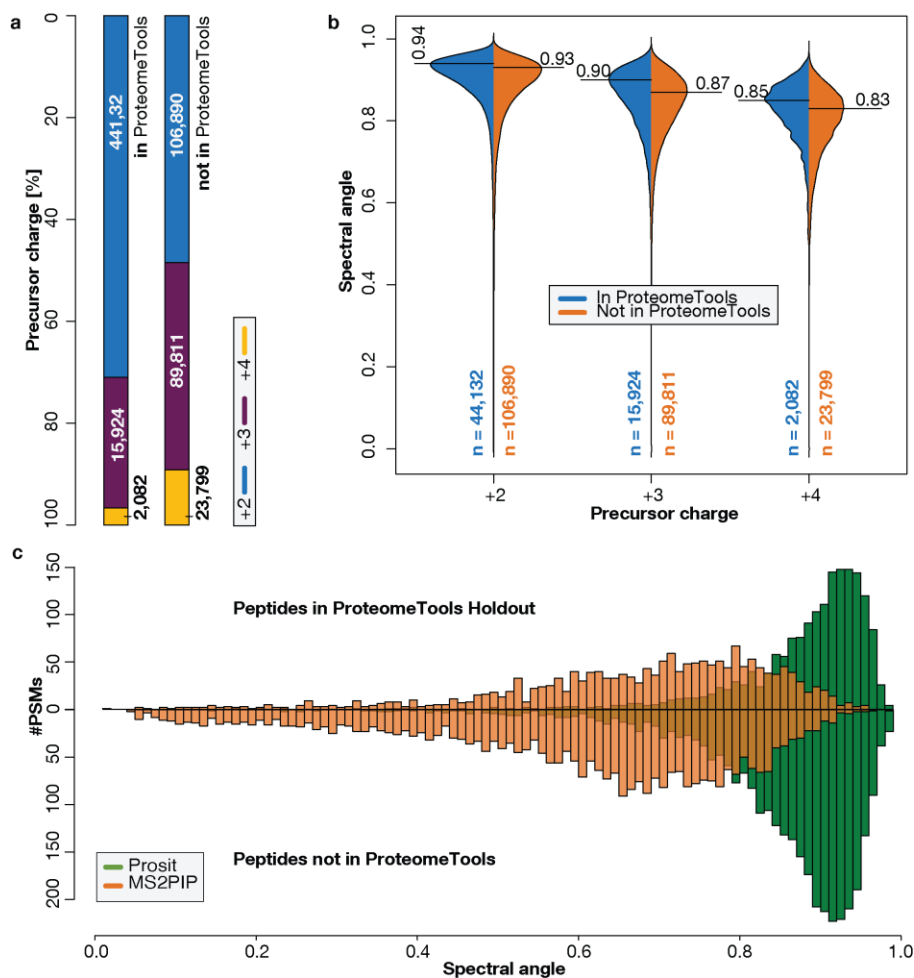
(d) Comparison of Prosit's and MS2PIP's fragment ion intensity prediction limited to spectra acquired at NCE 35 of the holdout set. Orange dots denote peptides that were (likely) part of MS2PIP's training data.



Supplementary Figure 5

Collision energy dependency of experimental and predicted spectra

Heatmap of the median spectral angle when comparing experimental vs experimental **(a)**, experimental vs predicted **(b)** and predicted vs predicted **(c)** spectra across 15 different normalized collision energies (NCEs) of ~40 synthetic peptides used for retention time and NCE calibration (Zolg et. al. 2017).



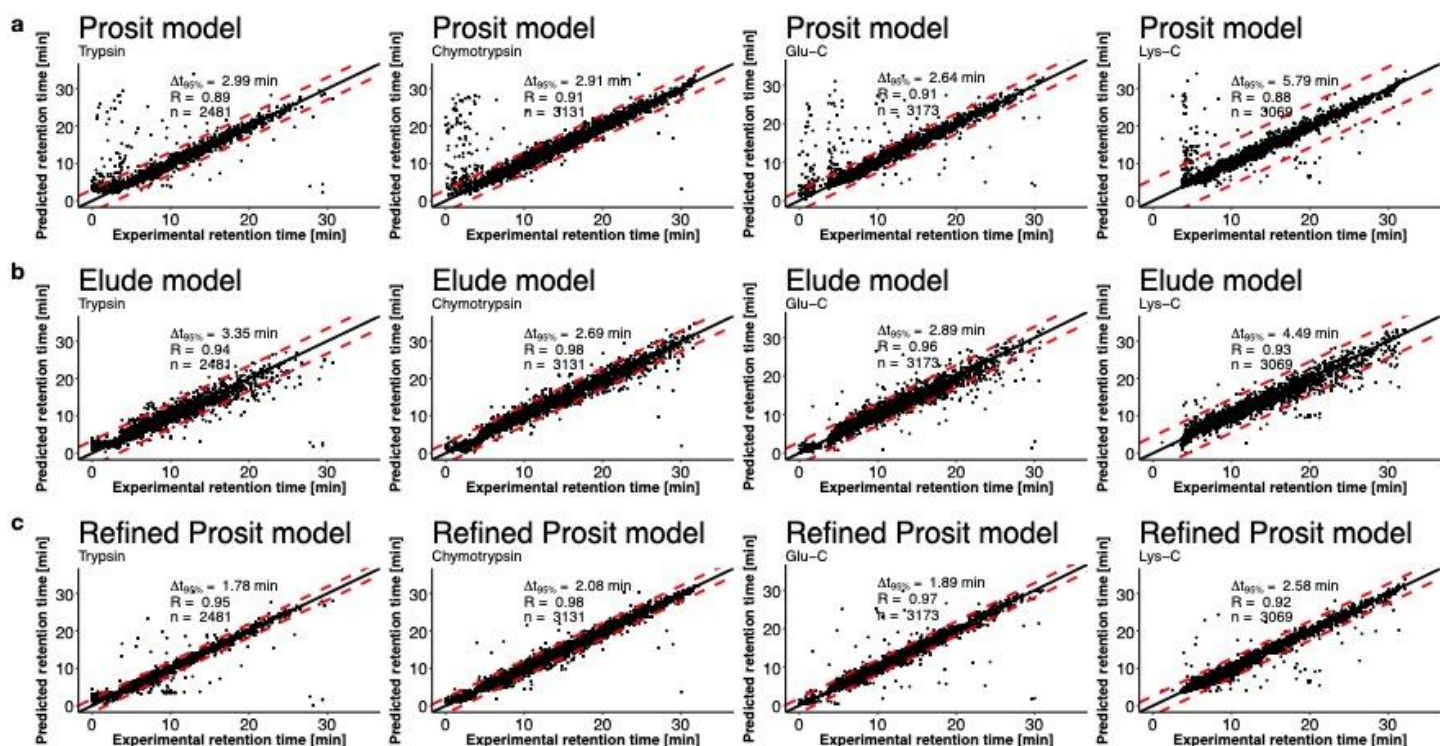
Supplementary Figure 6

Evaluation of model overfitting on internal and external datasets

(a) Comparison precursor charges of peptides from the Bekker-Jensen tryptic dataset. Peptides that were also part of the ProteomeTools Holdout dataset exhibit a different precursor charge distribution than those that were not.

(b) Spectral angle distributions by precursor charge for peptides in the Bekker-Jensen tryptic dataset split by whether they were also part of the ProteomeTools Holdout dataset.

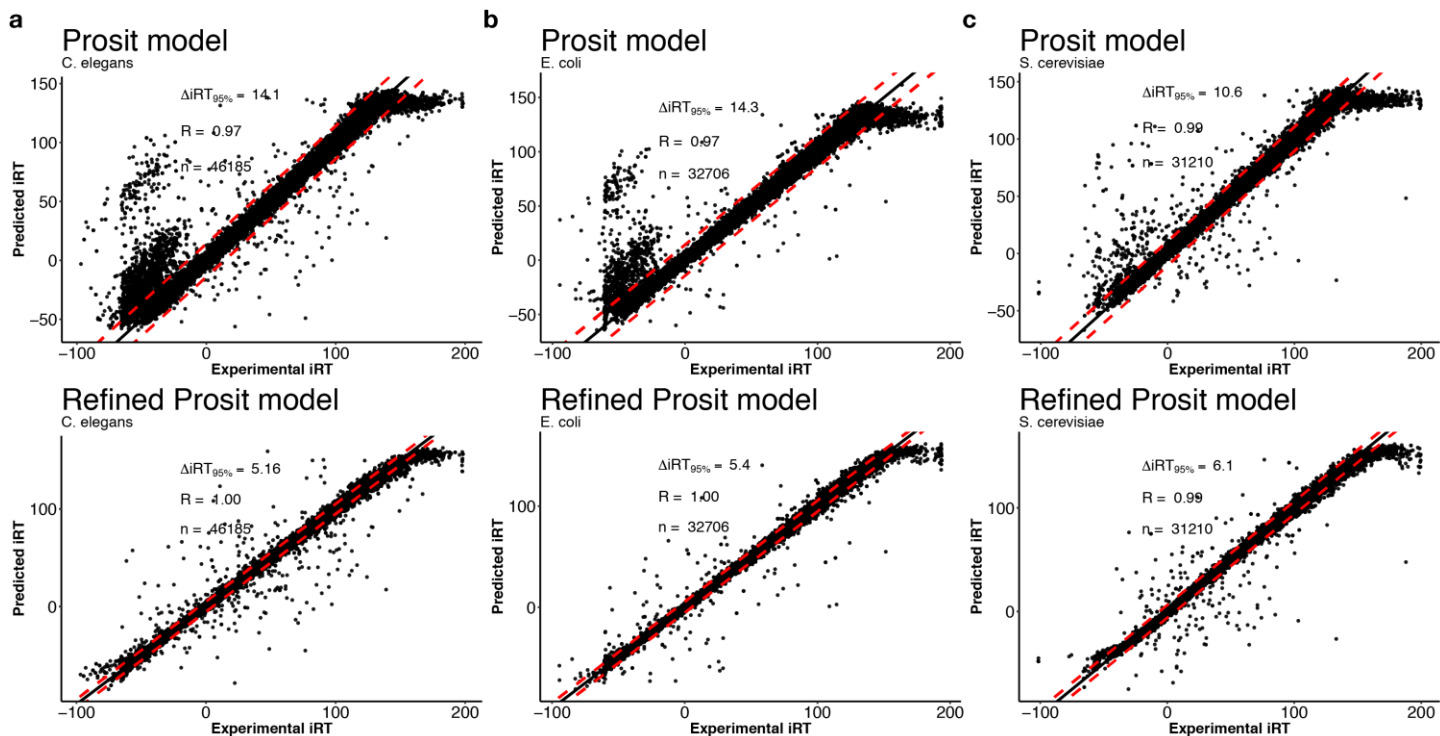
(c) Benchmark of Prosit's (green) and MS2PIP's (orange) fragment ion intensity prediction on tryptic peptides from the Bekker-Jensen dataset. The top histogram shows spectral angles for peptides that were also synthesized in the ProteomeTools project, but not used for training Prosit. The bottom histogram shows the distribution of spectral angles for peptides not part of ProteomeTools.



Supplementary Figure 7

Effect of the iRT model refinement for external datasets

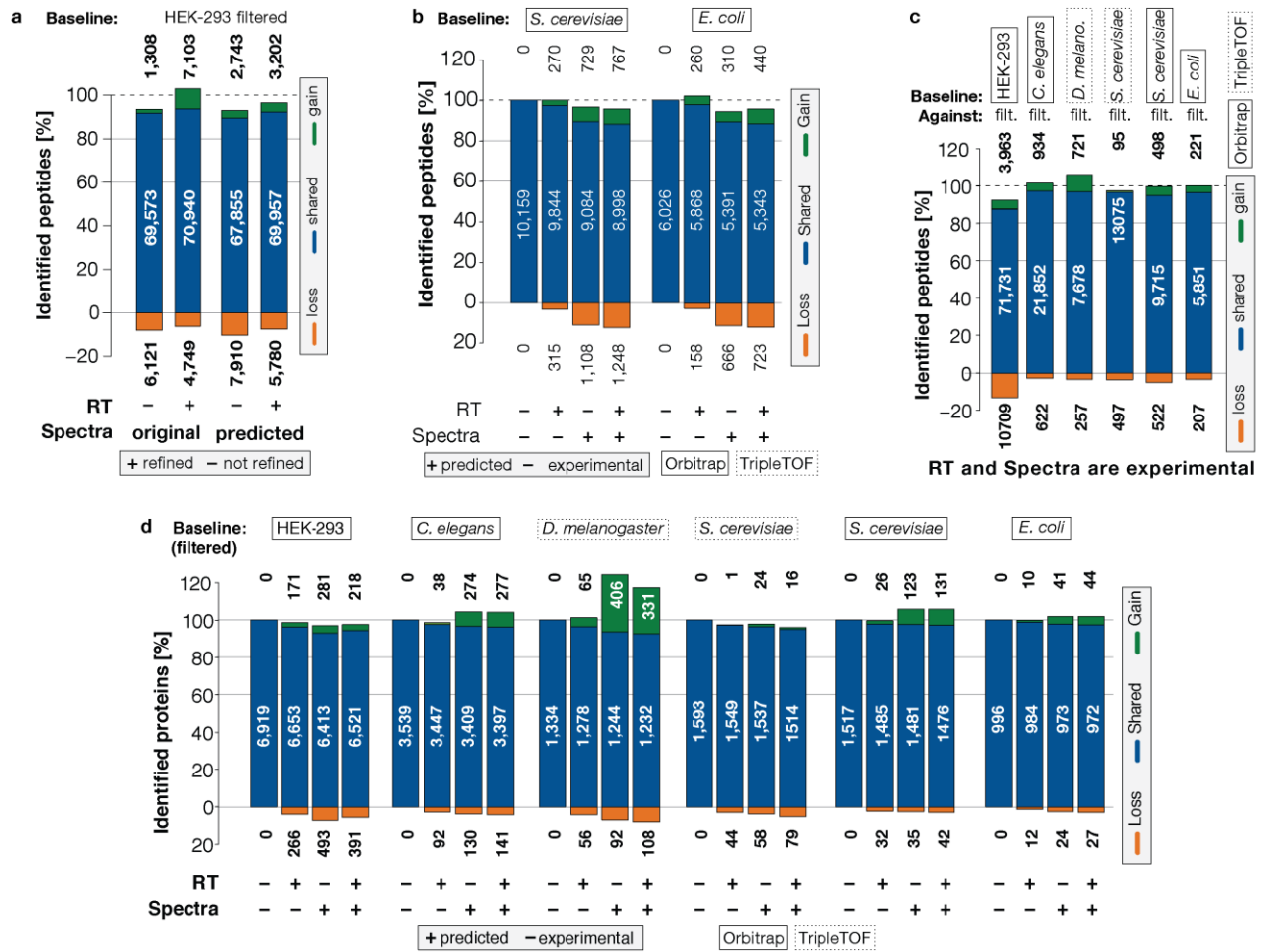
Predicted vs experimental retention times using the general Prosit indexed retention time (iRT) prediction model **(a)**, Elude **(b)**, or the refined Prosit model **(c)** on representative LC-MS/MS measurements for 4 proteases (left to right: Trypsin, Chymotrypsin, Glu-C and Lys-C) from the Bekker-Jensen et. al. dataset. Model refinement for Prosit was performed using the tryptic data from the same dataset. The required retention time window that would encompass 95% of all peptides is indicated by the red dashed lines. Sample number n and Pearson correlation are indicated.



Supplementary Figure 8

Effect of the refined Prosit iRT model on DIA spectral libraries

Evaluation of the general (top) and refined indexed retention time (iRT) (bottom) prediction model of Prosit on *C. elegans* (a), *E. coli* (b) and *S. cerevisiae* (c) data from the Bruderer et. al. dataset. For refinement, the project specific HEK library was used. The required iRT window that would encompass 95% of all peptides is indicated by the red dashed lines. Sample number n and Pearson correlation are indicated.



Supplementary Figure 9

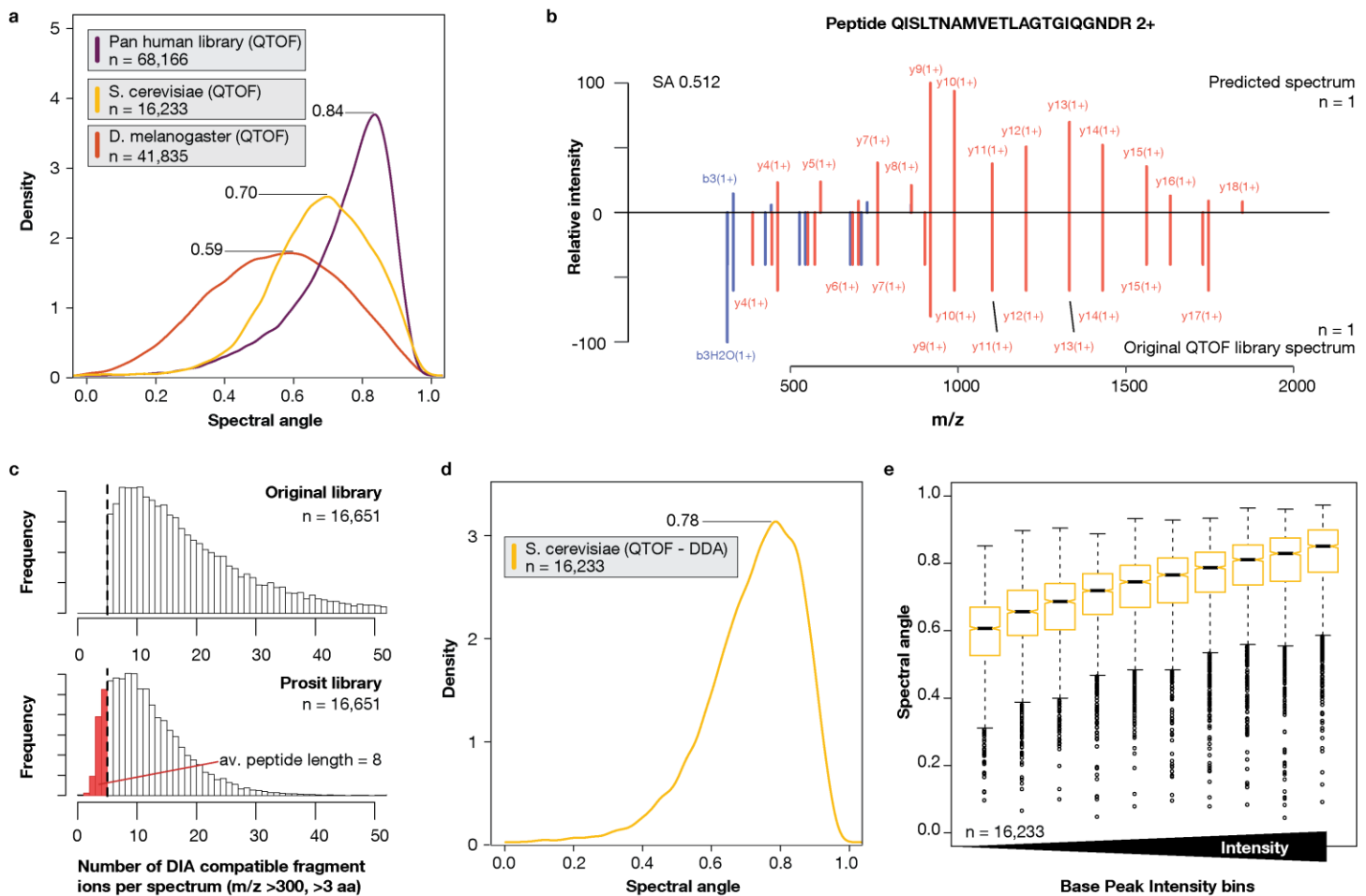
DIA analysis using predicted spectral libraries

(a) Impact of the retention time refinement using Prosit on the number identified peptides using either the general (indicated by "-") or refined (indicated by "+") Prosit indexed retention time (iRT) prediction model. The number of shared (blue), gained (green) and lost (orange) identified peptide sequences is plotted with respect to the original filtered library. iRT refinement was performed using the experimental retention time of the filtered HEK-293 data. See Supplementary Figure S8 for iRT model refinement analysis.

(b) Identical analysis as Figure 4 for *S. cerevisiae* and *E. coli*

(c) Re-analysis of Orbitrap/TOF based data independent acquisition (DIA)/SWATH datasets using predicted spectral libraries. Data and project specific spectral libraries were obtained from public repositories. To facilitate comparisons, the original library was filtered for entries that Prosit is not yet able to predict (other modifications besides oxidized methionine, neutral losses and peptides >30 amino acids). The original and filtered spectral libraries were queried against the DIA data using Spectronaut and the barcharts depict the number of shared (blue), gained (green) and lost (orange) identified peptide sequences when using the original filtered library compared to the original unfiltered library.

(d) Identical analysis as Figure 4, however protein-groups instead of peptides are displayed



Supplementary Figure 10

Comparison of predicted spectra with QTOF originated spectra

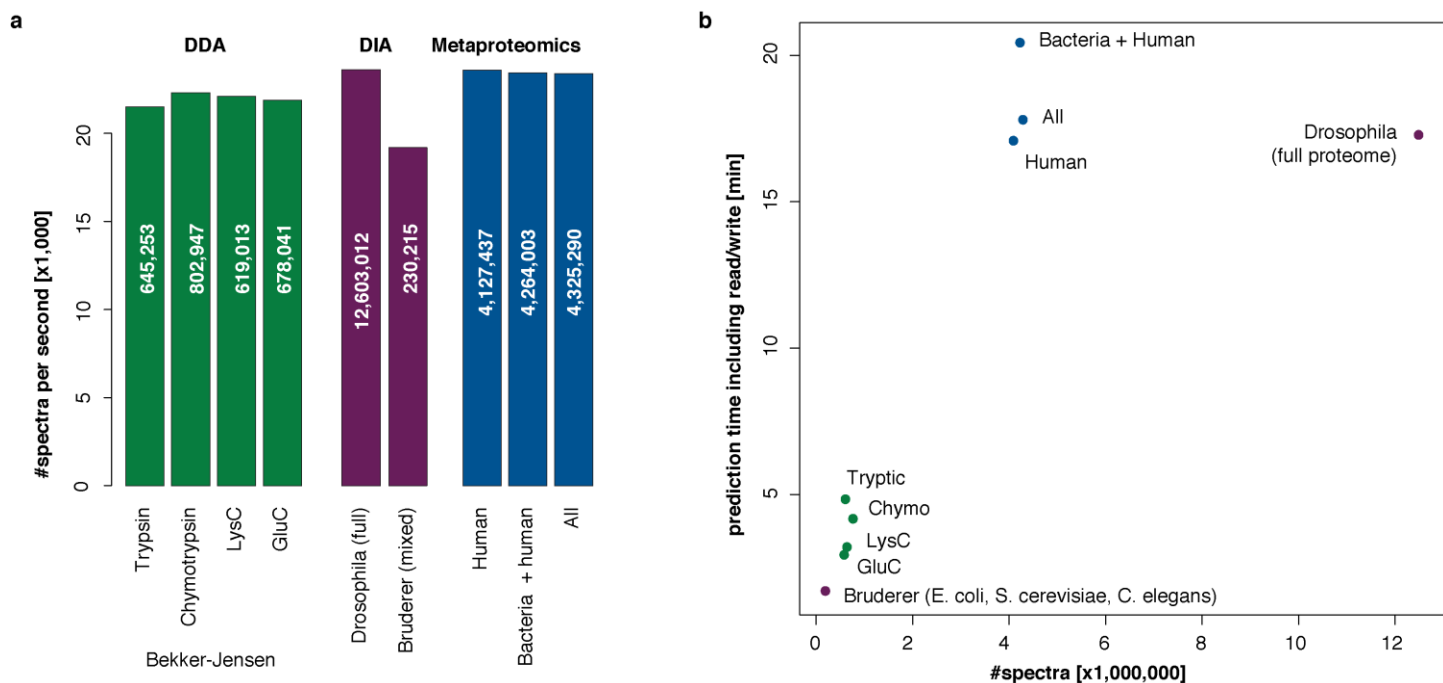
(a) Density distribution of normalized spectral contrast angles between predicted spectra and QTOF originated spectral libraries (Rosenberger et. al, Schubert et al, Fabre, et al). The spectral angle is calculated based on annotated fragment ions, excluding fragments with a neutral loss, less than 3 amino acids and m/z <300.

(b) Representative mirror spectrum of one predicted spectrum at normalized collision energy (NCE) 30 (top) vs one experimental spectrum contained in the *D. Melanogaster* QTOF library.

(c) Number of fragment ions annotated fragment ions, more than 3 amino acids and m/z >300 per spectrum in the *S. cerevisiae* library and after prediction.

(d) Density distribution of normalized spectral contrast angles between predicted spectra and DDA QTOF spectra for *S. cerevisiae* (Schubert et al.). Besides neutral loss fragments, all ions were accounted for.

(e) Density distribution of normalized spectral contrast angles between predicted spectra and DDA QTOF spectra for *S. cerevisiae* (Schubert et al.) as function of the most intense peak in the QTOF DDA spectrum.

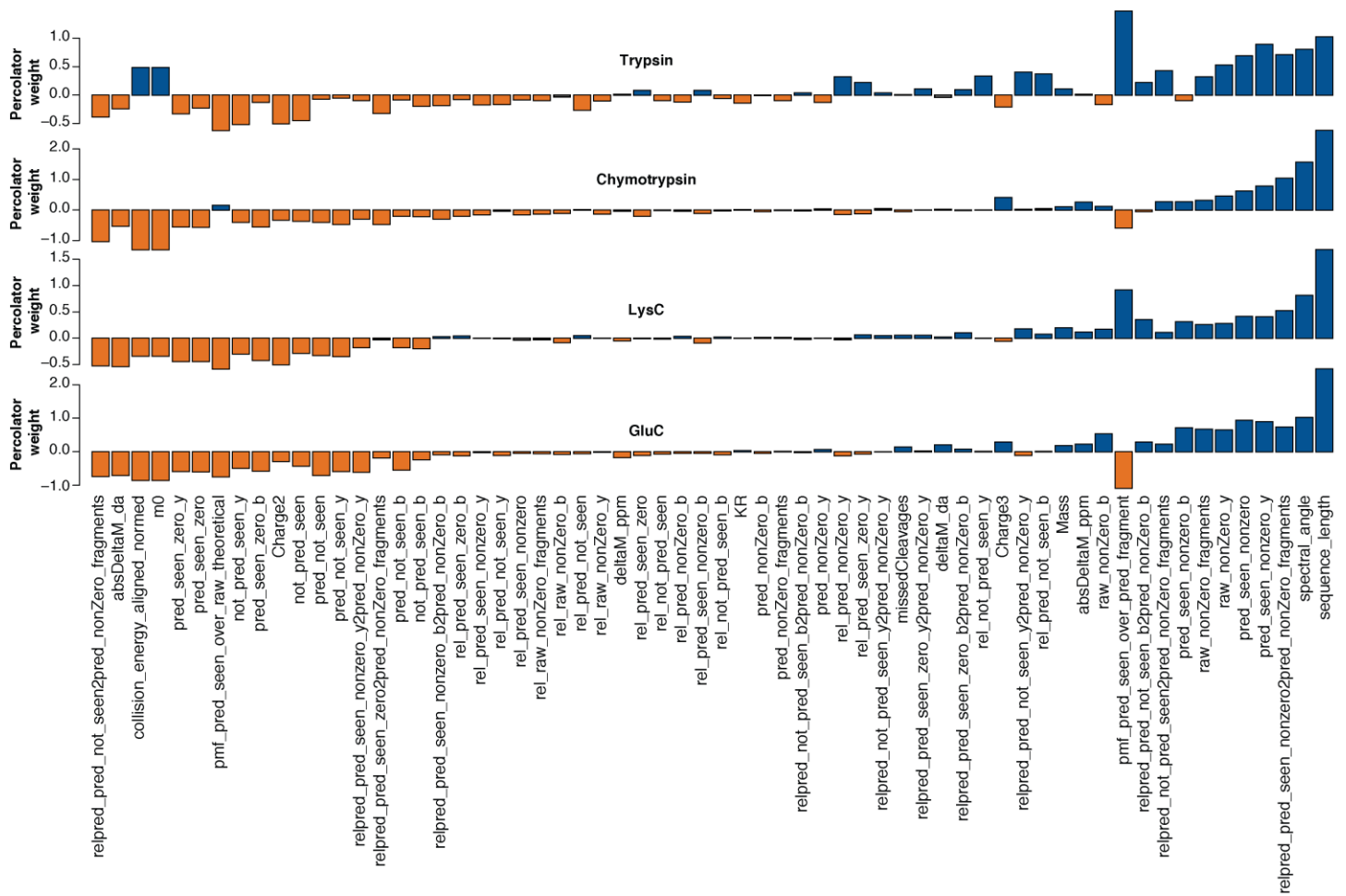


Supplementary Figure 11

Prediction performance analysis of Prosit

(a) Barplot of predicted spectra per second using Prosit's fragment ion intensity prediction across several datasets investigated in this study, excluding data transformation as well as read and write operations. Numbers in each bar indicate the total number of predicted spectra.

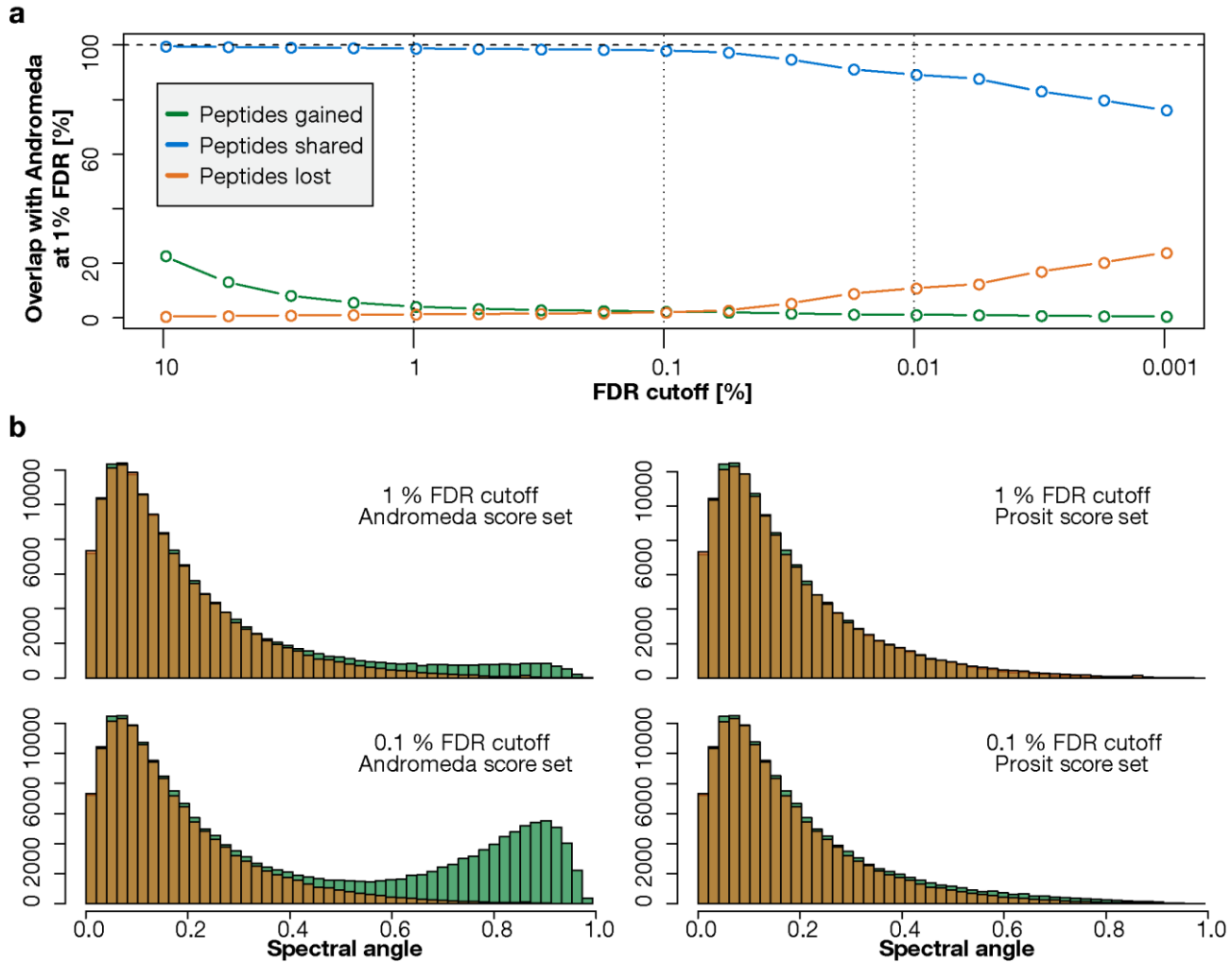
(b) Total prediction time including transformation, read and write operations plotted against the number of predicted spectra using Prosit's fragment ion intensity prediction model for differently sized datasets.



Supplementary Figure 12

Percolator feature weights

Barplots of final feature weights assigned by percolator for four different proteases when using the Prosit feature set (See Supplementary table 5 for description of the features). The evaluated percolator models were trained on Bekker-Jensen datasets with proteases (top to bottom): Trypsin, Chymotrypsin, Lys-C and Glu-C.

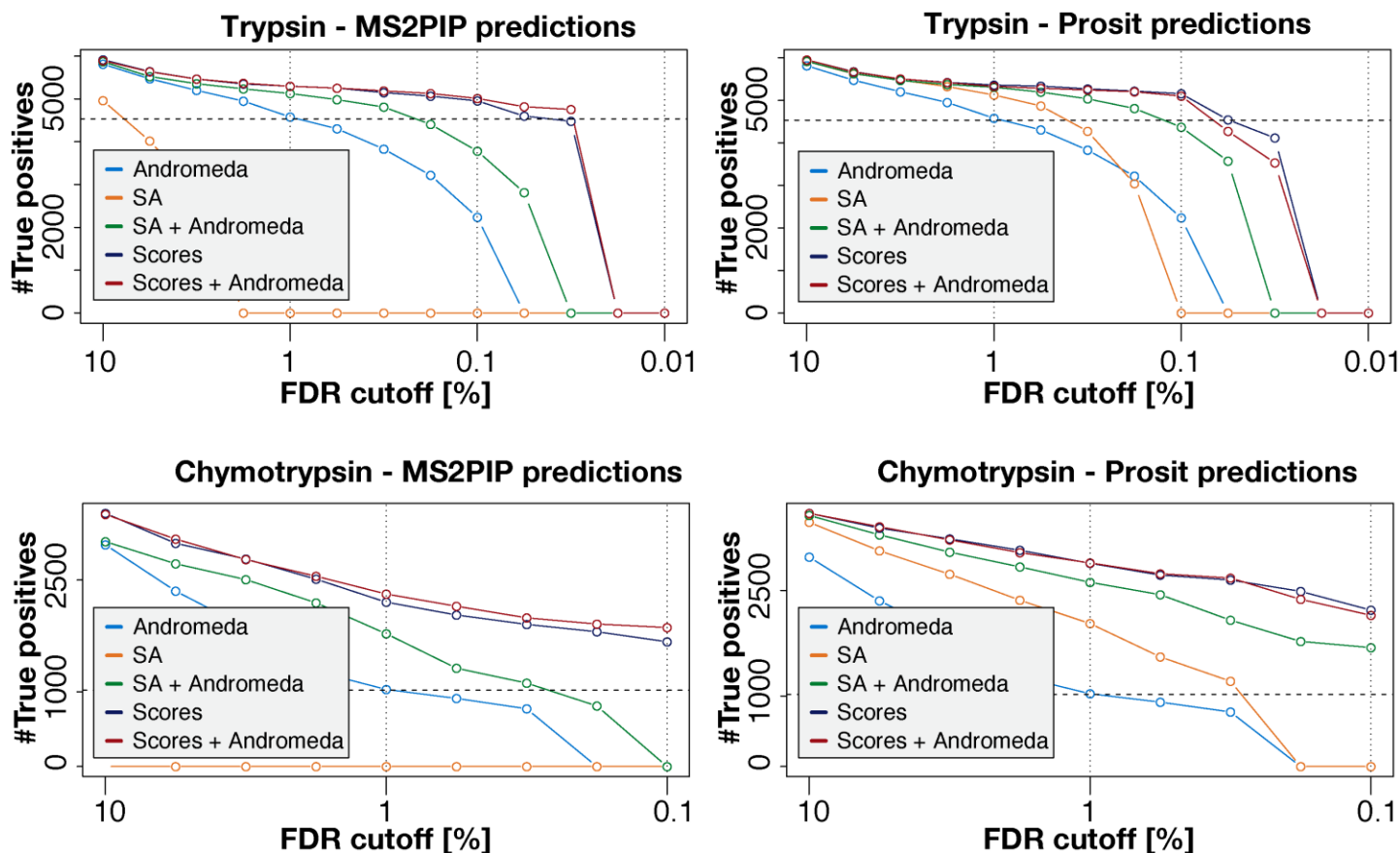


Supplementary Figure 13

FDR analysis Bekker-Jensen Trypsin

(a) Percent of shared (blue), gained (green) and lost (red) peptide identification when using the Prosit score set at different peptide level FDR cutoffs in comparison to the number of identification when using the Andromeda score set at 1% peptide level FDR.

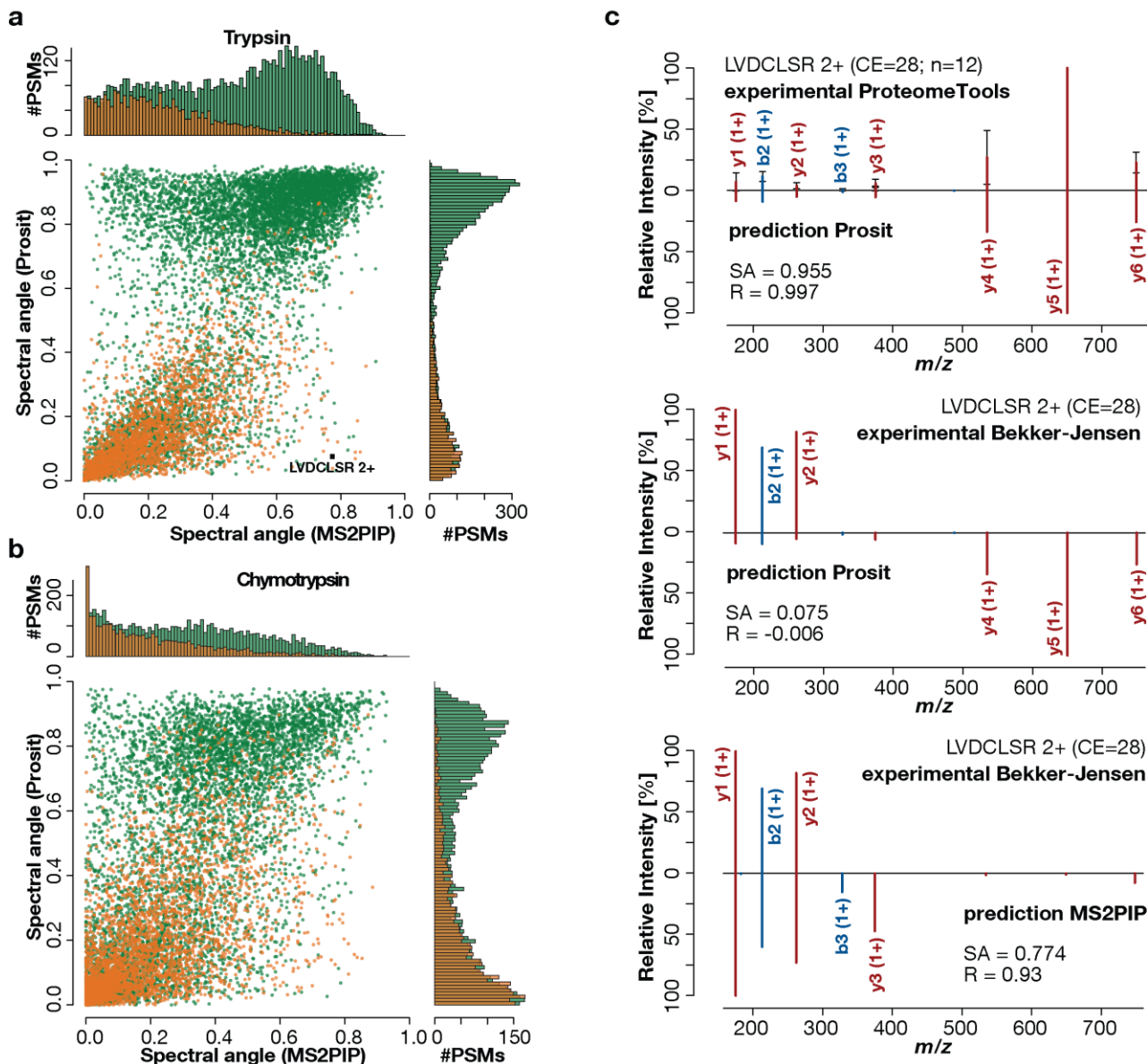
(b) Spectral angle distributions of decoy (orange) and false negative classified target (green) peptide spectrum matches (PSMs). The top panels are filtered at 1% peptide level FDR and the bottom panels are filtered at 0.1% peptide level FDR. The left panels show the distributions for the Andromeda and the right panel for the Prosit scores set.



Supplementary Figure 14

FDR comparison of Prosit and MS2PIP on Bekker-Jensen Trypsin and Bekker-Jense Chymotrypsin

Number of estimated true positive (#targets - #decoys at respective false discovery rate (FDR) cutoff) peptide spectrum matches using percolator at different peptide level FDR cutoffs when using the Andromeda (blue), full score set based on intensity predictions (orange) (see Supplementary Table 5 for feature set description). Dashed line indicates the number of true positive identifications when using the Andromeda feature set at 1% peptide level FDR. Top figures show the analysis for trypsin and bottom figures for chymotrypsin. On the left MS2PIP predictions were used and on the right Prosit predictions.



Supplementary Figure 15

Target-decoy analysis of Prosit and MS2PIP

(a) Comparison of the spectral angles of MS2PIP predictions to spectral angles of Prosit predictions for target (green) and decoy (orange) peptide spectrum matches generated by Andromeda. A random subset of 10,000 PSMs from the Bekker-Jensen tryptic dataset are shown. The PSM for peptide LVDCLSR is analysed in (c).

(b) As in (a), but on a random subset of 10,000 PSMS from the Bekker-Jensen chymotrypsin dataset.

(c) Analysis of a PSM for which Prosit's prediction has a lower spectral angle than MS2PIP's prediction. The PSM is highlighted in (a). Top: Prosit prediction compared to 12 experimental ProteomeTools spectra. Middle: Prosit prediction compared to the experimental spectrum from Bekker-Jensen. Bottom: MS2PIP prediction compared to the experimental spectrum from Bekker-Jensen. SA and R state spectral angle and Pearson correlation, respectively.

Supplementary Notes

Choice of the learning system

Representation learning and deep learning as one instance of it, is the idea to use data as input directly, rather than compressing the data to features and train models on these features. Before the advent of deep learning it was necessary to use features as lossy peptide representations and was as such commonly applied in Proteomics. Examples of this approach for fragmentation prediction are decision trees¹, MS2PIP², shallow neural networks³, boosting⁴, and for retention time prediction Elude⁵, SSCalc⁶. (regression). The recurrent neural network pDeep⁷ showed that deep learning can be applied to fragmentation prediction using a complete peptide sequence directly as input.

Previous models, such as pDeep, either trained separate models for different parameters, as for precursor charge, or ignored the influence of additional parameters, as for NCE. The wish to incorporate these parameters in a single model was a second factor for an encoder-decoder architecture. The latent space of this architecture gives the flexibility to add those parameters. Furthermore, the architecture gives flexibility in regard of the prediction problem. Prosit uses the same peptide sequence encoder architecture for iRT as well as fragmentation prediction.

The optimal number of layers, number of memory units in recurrent units, learning rate and dropout values were approximated empirically and chosen for feasibility on the given compute infrastructure. Specifically, more than one bi-directional layer did not improve the model compared to a uni-directional recurrent second layer. More than 2 recurrent layers did improve the model, but not so much as to justify the significant computational overhead. Long Short-Term Memory⁸ (LSTM) recurrent cells did not improve the model compared to Gated Recurrent Units⁹ (GRU), which are simpler computationally. The final architecture is described in the Methods section – Model Architecture (Figure S2).

Controlling the overfitting of Prosit's intensity prediction model

To ensure generalization of our model we applied three regularization techniques: early stopping¹⁰, Dropout¹¹ and the use of noisy data. For early stopping the data is split into Training (72%), Test (18%) and Holdout (10%) with peptide sequences only present in one of the three splits. The Training set is used to train the model and Test is used to control overfitting during training. We stop training if the Test error does not decrease for 10 epochs. Second, Dropout randomizes the error calculation at every training step by ignoring a portion of neurons in the network (50% in our case). As this renders a single neuron unreliable, the model's overall representation becomes distributed and more robust. Third, MS data is inherently noisy due to technical variation with fluctuating intensities and collision energies. Instead of using one consensus spectrum for a peptide, precursor charge and collision energy combination, we use the top 3 spectra ranked by Andromeda score resulting in around 36 spectra per peptide (assuming top 3 spectra for 2 precursor charges and 6 collision energies). Therefore, the model cannot memorize a perfect spectrum per combination, as it needs to minimize the overall error.

Figure S2b shows the model error for fragmentation prediction over training time for 5 random splits of the data. On an absolute scale (main panel), differences in model performance on Training, Test and Holdout are hardly noticeable. The inset zooms the spectral loss range of [0.08, 0.20], indicating only slightly lower error Training than on Test and Holdout. Notably, the error curves appear reproducible over all 5 splits, with min, max and mean errors over those splits being close in each training step.

Figure 2c shows that fragment intensity predictions for peptides in the Bekker-Jensen dataset that are included in ProteomeTools have higher spectral angles than those that are not in

ProteomeTools. One reason for this gap, is the different precursor charge distributions as shown in Figure S6a. The gap is especially severe for charge 3 precursors (Figure S6b). Charge 4 precursors have overall lower spectral angles, likely due to their rareness in the training data. Noteworthy, the difference of the apices of the spectral angle distributions of peptides being either present or absent in the ProteomeTools dataset is very similar across the 3 charge states, indicating that each group of spectra are subject to the same low level of overfitting, irrespective of the number of training data available.

Comparing Prosit's iRT prediction to SSRCalc. We compared Prosit to the retention time predictor SSRCalc (v Q.0)⁶ (Figure S3). As SSRCalc does not allow prediction of large datasets via its online service we selected a random subset of 10.000 peptides from our holdout dataset for comparison. Predicted Hydrophobicity Indexes were fitted with a linear model to experimentally determined indexed retention time (iRT) values and correlated.

iRT model refinement. Slight modifications to the gradient (i.e. isocratic start) and different physical setups of the LC (e.g. trap columns prior to analytical column) can impact the elution of certain classes of peptides. But, since most labs use at least the same mechanism of separation for online-LC, we tested transfer learning following the assumption that even small dataset (too small to learn deep learning-based iRT models from scratch) could be used to modulate the generic iRT model of Prosit and thus adjust the model to capture differences in online-LC which are not encoded in the ProteomeTools iRT values of peptides. For the Bekker-Jensen et al. dataset the tryptic DDA runs were first filtered for most intense identification per peptide sequence in every raw file and then filtered for Andromeda score >50. The refined model for HEK-293¹² was generated from 7000 randomly sampled entries of provided spectral libraries. For the re-analysis of *E. Coli*/*S. Cerevisiae*/*C. Elegans*¹² the whole *E. Coli* spectral library was used. Subsequently, the (indexed) retention times were scaled to be centered at 0 with a standard deviation of 1 (z-scoring). 80% of this dataset was used to refine the existing Prosit model and the remaining 20% were used to control for overfitting. Prosit was initialized with the model trained on the ProteomeTools data. iRT refinement took on average about 20 minutes and significantly increases lab-specificity (Figures S7 and S8).

For the Bekker-Jensen et al. dataset, we created classical iRT alignment models and Elude prediction models for every MS run in the Bekker-Jensen et al. dataset as we observed different gradient behaviour and length between MS runs.

Relating Pearson correlation and spectral angle. Using Pearson correlation (R) as a measure of spectrum similarity has been shown to be insensitive when spectra are very similar¹³. For example, R=0.99 and R=1.00 are very similar whereas for the same spectra the normalized spectral contrast angle (SA) (Methods) still captures a difference of 0.05 (SA=0.90 and 0.95 respectively). Figure S2c relates R with SA for all peptide sequence matches (PSMs) in our "Holdout" dataset, when comparing predicted to measured spectra. The region where spectra correlate strongly is particular important to achieve faster training convergence and better predictions.

Comparing Prosit's fragmentation prediction to MS2PIP and pDeep. In order to show the accuracy and generalization of Prosit, we compared our model to leading prediction models for fragment ion intensities, MS2PIP² and pDeep⁷. Both report considerably lower Pearson correlations for HCD spectra in their publications (pDeep R=0.90 overall, MS2PIP R=0.86 for +2 precursors, and Prosit R=0.99 overall). pDeep does not offer a prediction service and we were unable to run the available example code (<http://pfind.ict.ac.cn/download/pDeep.zip> downloaded 2017-11-20). An in-depth comparison is therefore limited to MS2PIP. Prosit shows better correlations for short peptides that have low precursor charges and missed cleavages (Figure S4a-b). Such biases are much more pronounced for MS2PIP and limit the applicability of such

predictions. For example, Prosit's correlations only slightly decrease from median SA=0.95 (R=1.00) for 7-mers to SA=0.90 (R=0.98) for 17-mers, whereas MS2PIP's correlations drop from SA=0.68 (R=0.85) to SA=0.5 (R=0.65) for the same peptides. These biases are likely related to unbalanced training data of MS2PIP as better correlations coincide with more training examples. Prosit takes into account the normalized collision energy (NCE) during training, effectively removing biases over five NCEs with an overall median SA of 0.91 (R=0.98). In contrast, MS2PIP achieves a reasonable median spectral angle of 0.7 (R=0.87) at NCE 35 but shows a bias at lower NCEs such as 20 (SA=0.4, R=0.52; Figure S4c). For the collision energy MS2PIP performs best (NCE=35), the similarities of Prosit's predictions beat similarities of MS2PIP's prediction in practically all cases (Figure S4d). When looking at predictions for an external dataset such as a subset of Bekker-Jensen Tryptic, Prosit outperforms MS2PIP by far (Figure S6c). Although Prosit's predictions are better for shorter peptides with fewer missed cleavages (i.e. such as those included in the ProteomeTools data), Figure S6c bottom shows that the same picture holds for peptides not included in ProteomeTools.

Figure S15 further details the advantage of Prosit's predictions over MS2PIP on targets as well as decoys. Prosit achieves higher correlations for target peptides and lower correlations for decoy peptides than MS2PIP resulting in a much stronger target decoy separation as shown in S15a for Trypsin and S15b for Chymotrypsin. Cases where MS2PIP achieves higher SA for targets than Prosit are often due to data rather than model quality. Figure S15 shows an example of such a case. ProteomeTools reference spectra match Prosit's predicted spectra very well (left), but do not match a spectrum from the Bekker-Jensen tryptic dataset for this peptide (middle). The higher correlation to the Bekker-Jensen spectrum by MS2PIP (right) is due to the overall low number of measured peaks. Figure S14 investigates the benefit of more accurate predictions for additional score generation and a subsequent FDR calculation by percolator. The baseline FDR calculation for both calculations is the *Andromeda* score set using Andromeda score and delta score as additional scores for percolator as described in Methods (processing of external data). We generated the same set of scores based on MS2PIP and Prosit as described for the *Prosit* score set in Methods. Figure S14 shows that Percolator benefits from scores based on MS2PIP's predictions compared to the *Andromeda* score although, the target decoy separation is not sharp (Figure S15). Focusing just on the SA score set, Percolator is not able to identify any targets at 1% FDR from MS2PIP predictions, whereas Prosit's SA score performs similarly to *Andromeda* score. (Figure S14). In the Chymotrypsin case Prosit was stronger relative to MS2PIP with all prediction-based score sets, indicating the influence of prediction quality on score quality for FDR calculation. The table below shows a 30% increase in identified PSMs when comparing Prosit-based percolator runs to MS2PIP-based runs on Chymotrypsin.

Table SN1: Comparing MS2PIP and Prosit by identified target PSMs

	Trypsin 1%	Trypsin 0.1%	Chymotrypsin 1%	Chymotrypsin 0.1%
Andromeda (Andr.)	4,579 PSMs	2236 PSMs	1,034 PSMs	0 PSMs
MS2PIP SA	0 PSMs	0 PSMs	0 PSMs	0 PSMs
MS2PIP SA + Andr.	5,127 PSMs	3,782 PSMs	1,781 PSMs	0 PSMs
MS2PIP Scores	5,298 PSMs	4,956 PSMs	2,205 PSMs	1,673 PSMs
MS2PIP Scores + Andr.	5,296 PSMs	5,013 PSMs	2,314 PSMs	1,864 PSMs
Prosit SA	5,125 PSMs	0 PSMs	2,030 PSMs	0 PSMs
Prosit SA + Andr.	5,305 PSMs	4,368 PSMs	2,615 PSMs	1,688 PSMs
Prosit Scores	5,358 PSMs	5,155 PSMs	2,886 PSMs	2,221 PSMs
Prosit Scores + Andr.	5,326 PSMs	5,099 PSMs	2,891 PSMs	2,147 PSMs

DIA spectral libraries

The ability to predict iRT and fragment ion intensity information for any peptide of interest enables the *in silico* generation of spectral libraries for data independent acquisition (DIA) analysis (Figure 4, Figure S9). We obtained the Orbitrap raw files for data re-analysis for human/*E. Coli*/*S. Cerevisiae*/*C. Elegans* from the PRIDE repository PXD005573¹². The spectral libraries on this repository were only available in binary, proprietary format, therefore the authors of the study generously provided the spectral libraries in textual format. To compare the actual prediction accuracy/quality the original library was filtered for modifications (besides oxidized methionine), peptides shorter than 7 and longer than 30 amino acids and neutral loss fragment ions. We acknowledge that biological samples do contain longer peptides and modified peptides and neutral loss ions and that the original library “as-is” will identify peptides Prosit is not yet able to predict. We are aware that the NCE determined for prediction (NCE=33) is only a proxy to match the DIA higher energy collision induced dissociation (HCD) spectra which are generally acquired using multiple (stepped) collision energies.

When comparing the spectral similarity distributions (Figure 4a) between the predictions and human/*E. Coli*/*S. Cerevisiae*/*C. Elegans* spectral libraries, it is apparent that they share overlapping distributions apexes near SA 0.9. The overall spectral similarity seems to be largely species independent and is high throughout.

In order to provide evidence that Prosit’s predicted retention time and spectral information are still well suited for DIA analysis, we base the analysis on peptides, ion types and modification states that are conceivable with the current version of Prosit. Therefore, the data shown in Figure 4b and Figure S9b and d is always in reference to the filtered original library. The comparison to the original spectral libraries can be found in Figure S9c. We also display additional bars for the two retention time prediction models used (Figure S9a).

We further investigated the transferability of Orbitrap based HCD scans to QTOF based HCD scans QTOF spectra exhibit a few characteristics that can complicate spectral comparison and the determination of fair measures to compare spectra: QTOF HCD spectra are usually aggregates of multiple TOF scans using a “rolling” collision energy that is ramped for individual

scans and aggregated into a spectrum, hereby averaging the HCD collision energies and increasing signal to noise. Ergo, spectra of low abundant species with very low signal to noise can exist if not enough scans were available for aggregation. Such spectra will result in lower spectral similarities compared to the predicted spectra, as the relative fragment ion intensities are not accurately represented anymore. However, high intense QTOF MS2 scans show a remarkable similarity to Orbitrap spectra as demonstrated using synthetic peptides¹⁴

To facilitate the comparison, we obtained two datasets and belonging spectral libraries from a *D. melanogaster*¹⁵ (ABSciex QTOF 6600, PRIDE identifier PXD006495) and a *S. cerevisiae*¹⁶ (ABSciex QTOF 5600, PRIDE identifier PXD001126) study as well as the QTOF originated Pan human spectral library (as downloadable in Spectronaut, ¹⁷). When comparing the three libraries with the predicted spectra (Figure S14a), that large differences between the libraries exist. The large Pan human spectral library exhibits a modus SA of 0.84 which is almost comparable to the Orbitrap based libraries investigated (Figure 4b) and more than expected, as Prosit's predictions were not specifically aligned to the HCD collision energy of the library. For the other two libraries the median and modus of the spectral similarity distribution is significantly lower (modus SA 0.70 for *S. cerevisiae* and modus SA 0.59 for *D. melanogaster*) despite collision energy alignment of the predictions. Investigation of this effect revealed that a sizable number of spectra in the *D. melanogaster* library suffer from a low signal to noise level. While the existence of fragment ions is in good agreement, the fragment ions in the QTOF spectrum do not reflect accurate relative intensities. Therefore, the calculated SA is low and proper ranking and selection of fragment ions for DIA data extraction is challenging. A representative scan which shows visible ion statistics and inaccurate ratios of fragment ion intensity is displayed in Supplemental figure Figure S10b. In addition, the spectra in the spectral library seem to be processed such that multiply charged fragment ion intensities are added to the respective singly charged fragment, changing the appearance of the spectrum. This explains the gain in identifications when using predicted spectra that exhibit proper signal to noise and represents ions in the charge state that are expected in the raw spectrum. Altogether, the spectral quality contained in the library explains the observed gain of 24% peptide identifications and 16% more protein identifications (Figure 4a, Figure S9c) when replacing the original spectra with high signal to noise predicted spectra with correct fragment ion charge state. These gains could be more extensive, did the predictions for short peptides not result in spectra containing less than the 6 required DIA compatible fragment ions (larger than 3 amino acids, larger 300 m/z and larger 5% base peak intensity – as defined during library building) which are lost for comparison in the current setup (Figure S10c).

To investigate, if the observed effects are study specific or originate from general TOF spectra characteristics, we used spectra from the *S. cerevisiae* study to directly compare DDA spectra with predicted spectra. In this comparison, all MaxQuant annotated peaks (besides neutral loss fragments) were compared to the predictions without employing a filtering process for mass range, number of amino acids in a fragment or minimum base peak intensity of fragment ions. The distribution Figure S10d displays a modus SA of 0.78, a value better than when comparing just the library due to more ions being factored in. However, the spectral similarity was still dependent on the dynamic range, hence signal-to-noise of the QTOF spectra (Figure S10e). While low signal spectra exhibit a rather poor median SA of about 0.6, spectra with lots of intensity and dynamic range performed identical with Orbitrap based scans. This demonstrates the good transferability of spectra to QTOF instruments. The results underline the importance of a well generated spectra library and suggest the replacement of low signal-to-noise spectra with consistent predictions to obtain consistent and confident results.

In conclusion, the impact of Prosit on DIA datasets is: First, and rather direct, predictions are precursor intensity independent and the obtained signal-to-noise level of the fragmentation spectra is consistently high for all peptides. Because of this, a homogenous spectral library can be generated. This is demonstrated on the Drosophila dataset, where the predicted library performs significantly better than the published library. Second, because of the ability to predict spectra at different collision energies, Prosit enables the calibration of spectral libraries and thus would allow the generation of consistent libraries for longitudinal studies (i.e. when collision energy settings drift) as well as its translation to other instruments (i.e. when collision energies settings or instruments change). Third, peptides observed in later stages of large projects (i.e.

longitudinal clinical study with additional DDA runs in between) can be added at any time to predicted spectral libraries without compromising its homogeneity. Fourth, assigning peptides to experimental spectra come with a risk that an actual spectrum of peptide A was falsely but confidently identified as peptide B. Storing this identification as peptide B in a spectral library will always give rise to a false positive identification and quantification in experimental data. This is, under normal circumstances, not the case for predicted spectra, as they will, most of the time, not look like another present peptide.

Currently, the use of comprehensive predicted spectral libraries is limited by the loss of sensitivity. Similar to DDA, additional features are required to enable more effective separation of true and random matches. Alternative approaches, like MSPLIT-DIA¹⁸, might be able to solve some of the issues by properly deconvoluting fragments and their intensity proportions, thus leading to less false positives, but we assume that the overall issue of the inflated search space would require some additional work. However, more research is necessary to address these issues comprehensively.

Advantage of prediction over large resources. ProteomicsDB currently stores ~43 million spectra covering ~700k peptides (including shared peptides). While this is in principle more than what is currently available in the ProteomeTools project, all these spectra are charge and isotope deconvoluted, thus their direct applicability is limited, because workflows operating on such data require pre-processed spectra as well. Even if ProteomicsDB would store raw MS2 spectra, acquisition parameters differ between labs. This heterogeneity would require significantly more effort to build homogenous spectral libraries from data collected in ProteomicsDB or otherwise require the deconvolution of mixed similarity distributions introduced by spectra matched from different labs, e.g. collision energies. In contrast, Prosit allows the prediction of raw (meaning unprocessed) intensities and thus enables the direct comparison to acquired raw spectra (from any acquisition method) without the requirement of any additional MS data pre-processing.

In addition, spectral libraries generated with data from ProteomicsDB would come with the same limits as any other experimental spectra library, namely its inability to identify peptides which were previously not seen (e.g. unobserved proteins, other proteases). Much more importantly in this context is the question of how to model random (decoy) events. Without proper decoys, FDR estimation rely on heuristic approaches to generate decoy spectra. This is a known challenge in spectral library searching with no generally accepted solution yet¹⁹. Prosit, on the other side, can predict spectra for decoy peptides, overcoming the issue of potentially generating “false” decoy spectra entirely and thus allows accurate FDR estimation. This is affirmed by the near perfect overlap between decoy and false positive target peptide spectrum matches in the lower spectral angle region as shown in Figure 5a. The fact that these distributions overlap nearly perfectly suggest correct decoy generation and allows proper estimation of the number of falsely matched spectra.

Proteotypicity prediction and peptide selection. Selecting appropriate peptides for MRM/PRM assays or the content of a spectral library are essential for DIA datasets. With Prosit, a possible selection process for targeted assays is the prediction of a set of possible peptides (which comes at no cost and little effort), which can be used for an initial targeted proteomics experiment. Subsequently, positive matches (hits) can be synthesized as standards and for validation. This procedure was successfully employed for in-house experiments as well as collaborative projects and led to the successful quantification of a so far unobserved protein (data not shown).

In addition, while large resources can be used for selecting appropriate peptides, we have strong evidence that the entire sample preparation workflow performed in a lab has a significant influence on the presence and absence of peptides, despite the protein being available²⁰. This is apparent when comparing proteotypic peptides across different labs. We have trained a proteotypicity model on data downloaded from ProteomicsDB and tested this model on experimental data (data not shown). While the model showed an AUC of ~0.95 on a hold-out dataset provided by ProteomicsDB, the overall prediction quality on external datasets was

significantly worse, likely because of different sample preparation protocols (differences in e.g. digestions, desalting or offline separation). This was even more apparent when training and evaluating predictors for the pre-dominant precursor charge state of peptides (modulated by e.g. different solvents in LC). We concluded that while fragmentation is largely conserved between different instruments, adjustable by modifying the NCE, other properties, such as retention time, proteotypicity and dominant charge state (in this order from good to worse), are significantly harder to predict using a generic model.

Prediction speed. Inference speed is constrained by read and write operations. Figure S11a shows the speed of only the prediction across datasets for the fragment intensity model with up to 1.4 M spectra per minute. Fig S11b shows the time for prediction and read/write operations in relation to the number of spectra. A prediction of 12.6 M spectra requires ~17 min. The iRT prediction of 12.6 M peptide sequences requires ~ 8 min. We use a server with 1 Nvidia Titan Xp GPU and 506 GB RAM for prediction.

We have made extensive use of the HDF5 format for data handling, because binary formats in general provide efficient read/write access and thus reading/writing data comes with minimal overhead. Because of the lack of a common spectral library format ²¹, we allow the export of prediction in two commonly used, textual library formats. However, writing textual files comes with significant overhead and depending on the format, information such as the peptide sequence are stored very redundantly. This renders data transformation and export to (third party) usable library formats the most time-consuming part of the prediction process.

Percolator score set generation. When only using the spectral angle to discriminate true from false matches, the overall separation by percolator is not as good as when using the Andromeda score (Figure S13). One reason for this is that the spectral angle does not score the number of observed matching fragment ions with respect to the number of all theoretical fragment ions, which leads to several decoy PSMs which achieve high spectral angles of >0.9 while Andromeda scores for decoys max out at around 150. Adding additional features such as the number of observed fragments removes this problem. Therefore, we built additional scores build on Prosit's predictions to balance out this problem. Exploring the effect of the generated scores is possible because percolator weights features based on their impact for target-decoy separation (Figure S12). As expected, many of the most important scores cannot be generated without predictions of experimental quality.

False discovery rate cutoff analysis. A comparison of the distribution of decoy PSMs to the distribution of false positive target PSMs (i.e. >1% FDR) gives some insight into the differences of the separation capabilities of Andromeda and Prosit²². When ranked by Andromeda the proportion of target sequences not making the FDR cutoff is significant at 1% and drastically increases with lower peptide FDR levels as seen in the left column of Figure S13 for the Bekker-Jensen Tryptic dataset. When ranked by Prosit, there are virtually no false negatives for FDR levels as low as 0.1% (right column). Only at 0.01% false positives can be seen. This exemplifies a broader pattern seen as example in Figure 5d: correlating experimental spectra to predictions dramatically helps percolator to correctly separate target and decoys. Similarly, this holds for true positives, too. With Prosit, the separation of true from false positive matches is much better. The figures below show FDR cutoffs for Chymotrypsin (Figure SN1), Glu-C (Figure SN2) Lys-C (Figure SN3) and for the metaproteomics analysis of Human (Figure SN4), Human + Bac (Figure SN5), All (Figure SN6) and All + IGC (Figure SN7)

Figure SN1: FDR analysis Bekker-Jensen Chymotrypsin

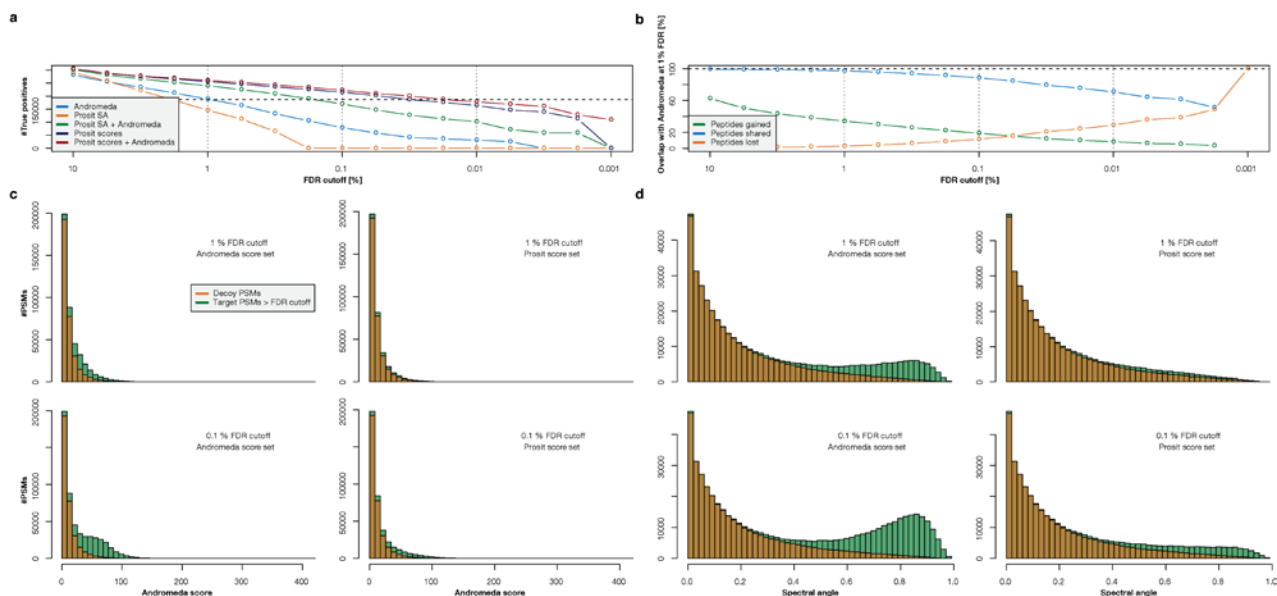


Figure SN2: FDR analysis Bekker-Jensen Glu-C

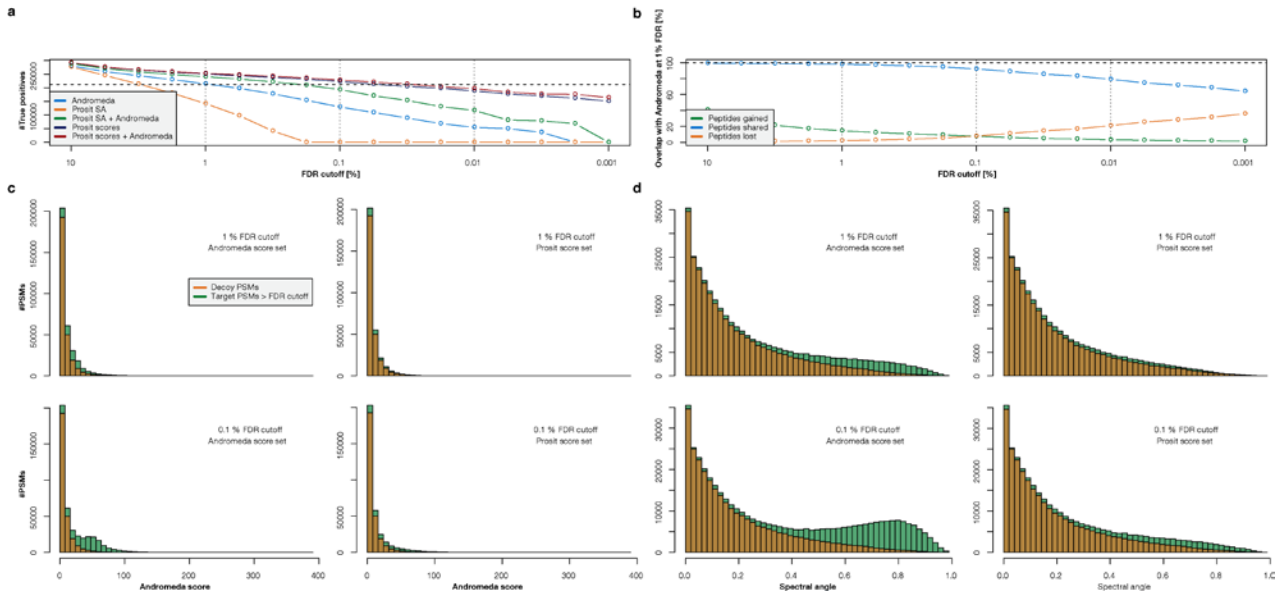


Figure SN3: FDR analysis Bekker-Jensen Lys-C

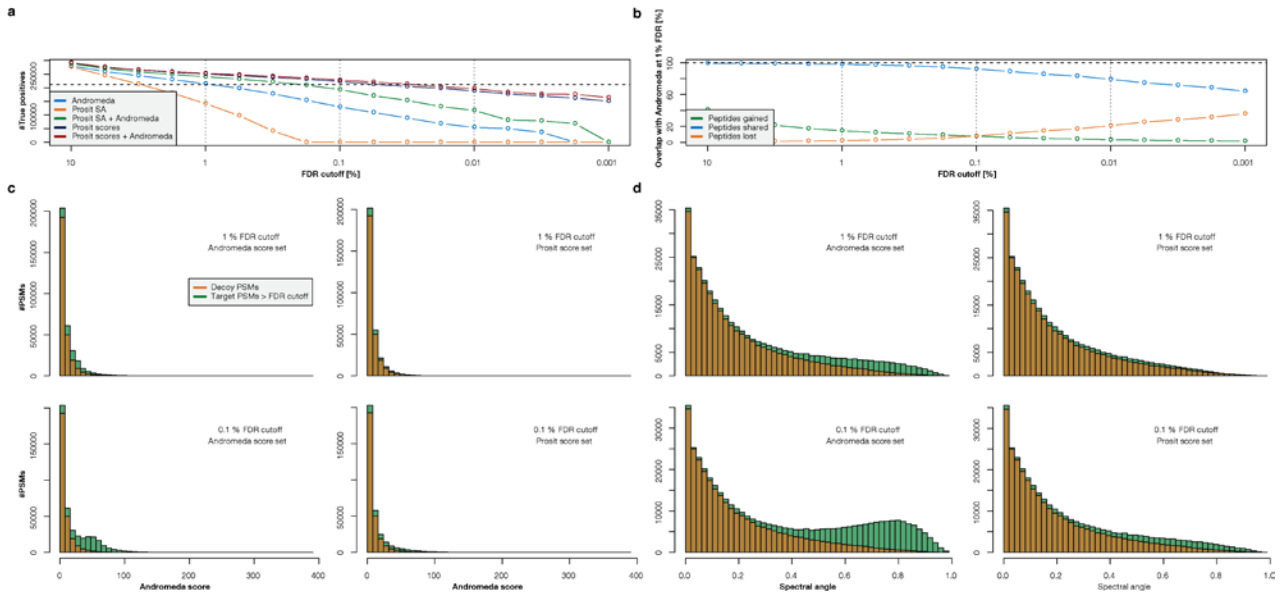


Figure SN4: FDR analysis metaproteomics using the Swissprot Human database

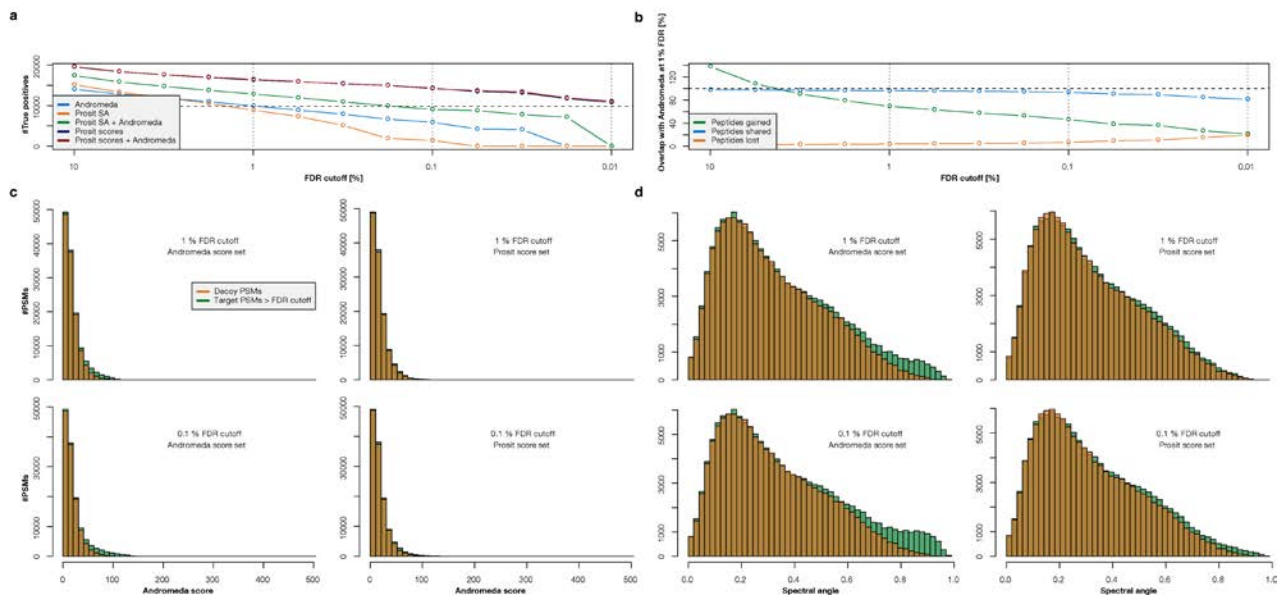


Figure SN5: FDR analysis metaproteomics using the Swissprot Human + Bacteria database

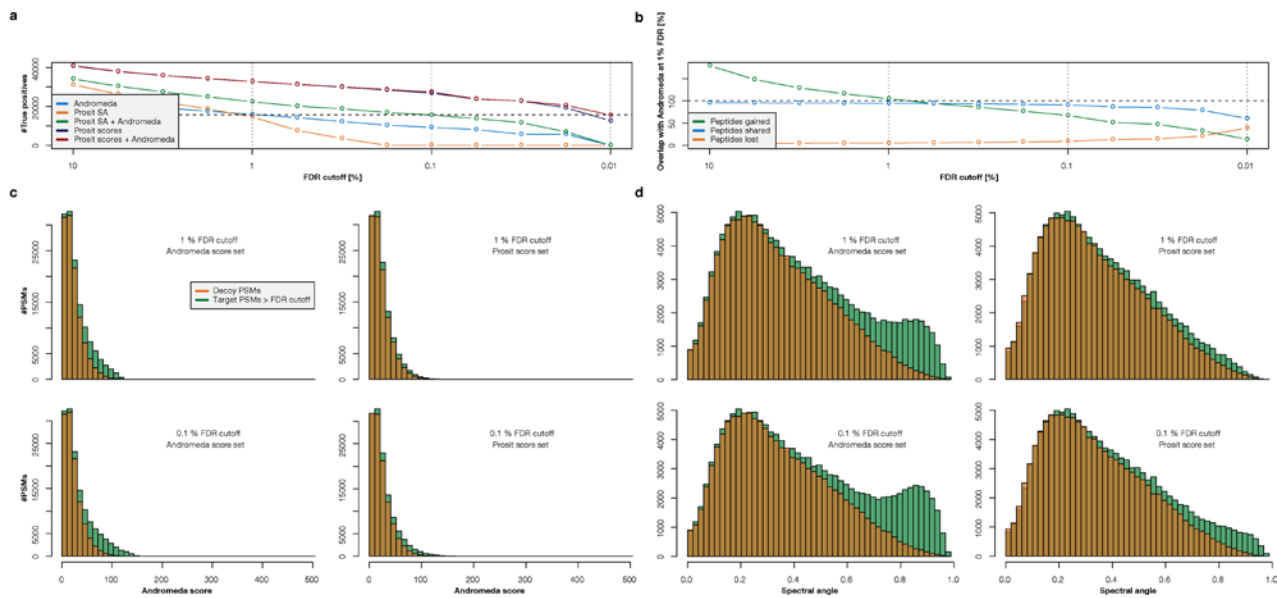
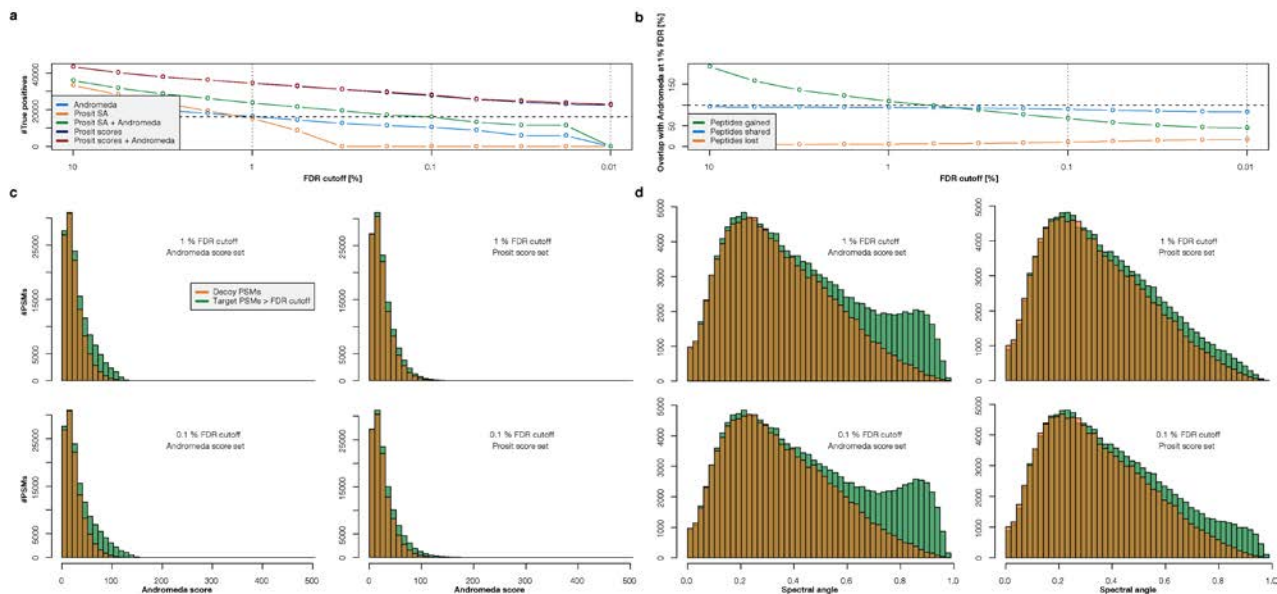
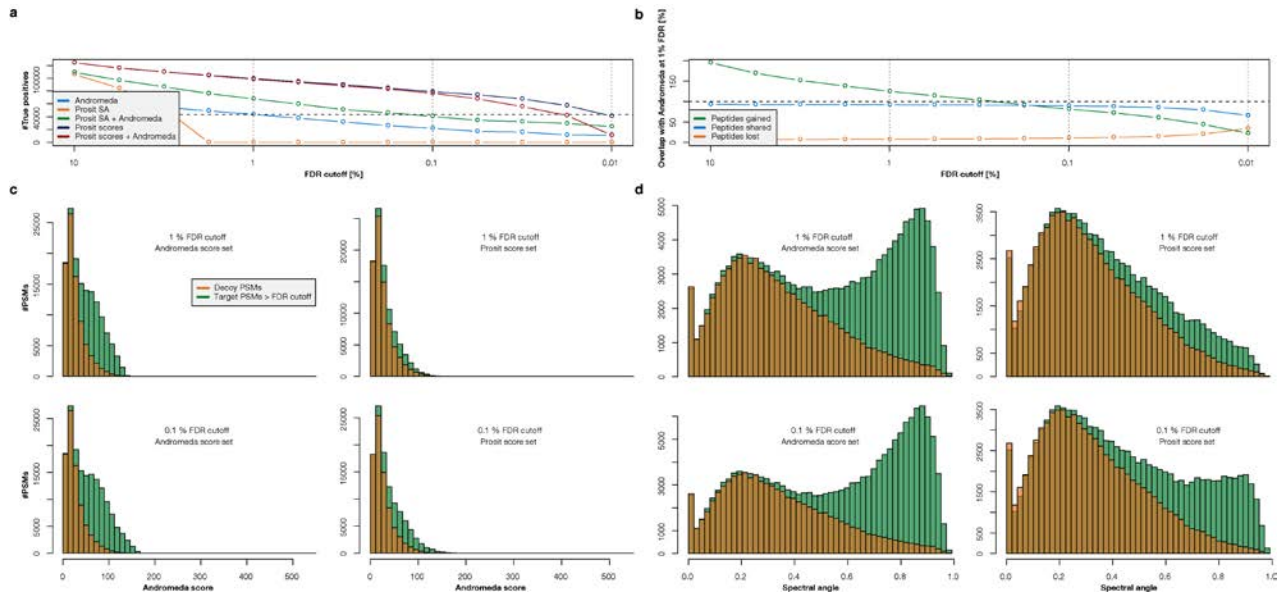


Figure SN6: FDR analysis metaproteomics using the Swissprot All database



(a) Number of estimated true positive ($\# \text{targets} - \# \text{decoys}$ at respective false discovery rate (FDR) cutoff) peptide spectrum matches using percolator at different peptide level FDR cutoffs when using the Andromeda (blue), spectral angle (SA; orange), Andromeda + spectral angle (green), Prosit (blue) and Prosit + Andromeda (violet) feature set (see Supplementary Table 5 for feature set description). Dashed line indicates the number of true positive identifications when using the Andromeda feature set at 1% peptide level FDR. **(b)** Percent of shared (blue), gained (green) and lost (red) peptide identification when using the Prosit feature set at different peptide level FDR cutoffs in comparison to the number of identification when using the Andromeda feature set at 1% peptide level FDR. **(c)** Andromeda score and **(d)** spectral angle distributions of decoy (orange) and false negative classified target (green) peptide spectrum matches (PSMs). The top panels are filtered at 1% peptide level FDR and the bottom panels are filtered at 0.1% peptide level FDR. The left panels show the distributions for the Andromeda and the right panel for the Prosit feature set.

Figure SN7: FDR analysis metaproteomics using the IGC + Swissprot All database

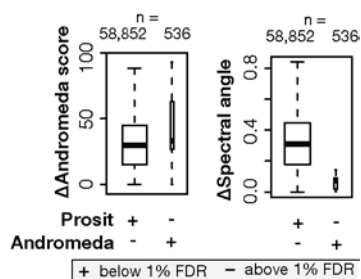


(a) Number of estimated true positive ($\# \text{targets} - \# \text{decoys}$ at respective false discovery rate (FDR) cutoff) peptide spectrum matches using percolator at different peptide level FDR cutoffs when using the Andromeda (blue), spectral angle (SA; orange), Andromeda + spectral angle (green), Prosit (blue) and Prosit + Andromeda (violet) feature set (see Supplementary Table 5 for feature set description). Dashed line indicates the number of true positive identifications when using the Andromeda feature set at 1% peptide level FDR. **(b)** Percent of shared (blue), gained (green) and lost (red) peptide identification when using the Prosit feature set at different peptide level FDR cutoffs in comparison to the number of identification when using the Andromeda feature set at 1% peptide level FDR. **(c)** Andromeda score and **(d)** spectral angle distributions of decoy (orange) and false negative classified target (green) peptide spectrum matches (PSMs). The top panels are filtered at 1% peptide level FDR and the bottom panels are filtered at 0.1% peptide level FDR. The left panels show the distributions for the Andromeda and the right panel for the Prosit feature set.

Delta score analysis for two top-ranked PSMs. When up to 15 PSM candidates were included in the metaproteomics analysis Andromeda only rarely identified peptides not identified by Prosit (Figure 6c bottom right corner). In contrast, Prosit identified many MS/MS spectra that did not

pass the Andromeda's scoring threshold (top left corner). To investigate the differences in more detail, we calculated the Andromeda delta score and delta spectral angle between the top and second-ranked PSM for Andromeda and Prosit respectively and plotted the distributions for spectra where Andromeda and Prosit disagreed (Figure SN8). Based on Andromeda score (left panel), there was no difference in median delta scores for PSMs accepted by Andromeda but rejected by Prosit or vice versa. Conversely, a median delta SA of 0.3 for PSMs accepted by Prosit but rejected by Andromeda and a delta SA of near zero for PSMs rejected by Prosit but accepted by Andromeda (right panel) indicated that Prosit is more confident in its top-ranking PSM.

Figure SN8: Delta score analysis for the two top-ranked PSMs.



Delta scores for top ranked to second-best ranked PSMs when ordered by spectral angle (Prosit, bottom) or Andromeda score (top panel). The two boxes on the left are confident identifications by the Prosit set of scores but not by the Andromeda set of scores. The two boxes on the right show the opposite. The boxes are drawn to scale (number of PSMs) and indicate the interquartile range (IQR) and its whiskers represent 1.5*IQR values. The median is indicated, and outliers are not shown. This shows that PSMs that are ambiguous in Andromeda can often be distinguished by Prosit.

Data availability. The data made available on three different platforms, depending on the nature of the data. Proteomics data, including Raw MS data, search results and percolator results are made available on PRIDE. Code to train and use the fragmentation and iRT models is shared on github.com. All other data is shared on figshare.com. This includes description and weight files of the trained models, and full versions of the supplemental tables

ProteomeTools raw MS data

Synthetic peptide spectra are available at <https://www.proteomicsdb.org> and updates to the resource are available at www.proteometools.org. The mass spectrometric data of the ProteomeTools synthetic peptide sets have been deposited with the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD010595.

The raw file naming convention is the following:

`<PlateID>_<WellID>-<Set>_<Pool>_<SynthesisReplicate>_<Aliquot>-<Measurement>-<Gradient>-<MeasurementReplicate>.raw`

Example: 02079a_BA1-TUM_isoform_1_01_01-DDA-1h-R3.raw

Internal *PlateID* is 02079a, internal *WellID* BA1, *set* is TUM_isoform, *pool number* 1, first *synthesis replicate*, first *aliquot*, *measurement* method was data dependent survey run, 1h LC *gradient*, *third measurement replicate*.

The peptide set is either the “isoform” set (TUM_isoform), “Missing gene addon” (TUM_second_addon), “TMT proteotypic” set (TUM_proteo_TMT) or “Re-synthesis proteotypic” (TUM_missing_in_first).

Measurement method is either the survey run (DDA), HCD run (3xHCD), IonTrap run (2xIT_2xHCD) or ETD run (ETD).

Re-analysis of external data

For the benchmarks against external datasets (Figure 2, 3, 5 and 6), we provide the MaxQuant search results of the external dataset, extracted raw scans, Prosit's prediction and Percolator

output. For the reanalysis of DIA data using predicted spectral libraries (Figure 4), we provide the full Spectronaut search, search report and spectral libraries used. For the improved identification in metaproteomic samples we provide generated raw files, MaxQuant search results, extracted raw scans, Prosit's prediction and Percolator output.

Data and analyses described in this manuscript have been deposited with the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the separate dataset identifier PXD010871.

Trained model availability

Model definition YAML files and Model weights (HDF5 files) for both fragmentation and retention time prediction are available at figshare.com/projects/Prosit/35582/. This repository includes datasets from ProteomeTools we used as "Training", "Test" and "Holdout" datasets in HDF5 format.

Code availability

Code for training, prediction and to run a server on your local hardware is available at www.github.com/gessulat/prosit/. The code can be used with the pre-trained models made available on figshare.

Supplementary Tables

The supplementary tables available with the manuscript are filtered version due to space limitations. You can find the full versions at figshare.com/projects/Prosit/35582/.

References

1. 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. *Nature biotechnology* **22**, 214-219 (2004).
2. Degroeve, S., Maddelein, D. & Martens, L. MS2PIP prediction server: compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research* **43** (2015).
3. Arnold, R.J., Jayasankar, N., Aggarwal, D., Tang, H. & Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pac Symp Biocomput*, 219-230 (2006).
4. Frank, A.M. Predicting Intensity Ranks of Peptide Fragment Ions. *Journal of Proteome Research* **8** (2009).
5. Moruz, L., Tomazela, D. & Käll, L. Training, selection, and robust calibration of retention time models for targeted proteomics. *Journal of Proteome Research* **9** (2010).
6. 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. *Analytical Chemistry* **78** (2006).
7. Zhou, X.-X. et al. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Analytical Chemistry* **89**, 12690-12697 (2017).
8. Hochreiter, S. & computation, S.-J. Long short-term memory. *Neural computation* (1997).
9. Cho, K. et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
10. Caruana, R., Lawrence, S. & Giles, L.C. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems* (2001).

11. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning* **15** (2014).
12. Bruderer, R. et al. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Molecular & Cellular Proteomics* (2017).
13. Toprak, U.H. et al. Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics. *Molecular & Cellular Proteomics* **13**, 2056-2071 (2014).
14. Zolg, D.P. et al. Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods* **14**, 259-262 (2017).
15. Fabre, B. et al. Spectral Libraries for SWATH and *Solanum lycopersicum*. *PROTEOMICS* **17** (2017). -MS Assays f
16. Schubert, O.T. et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols* **10**, 426-441 (2015).
17. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data* **1** (2014).
18. Wang, J. et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nature Methods* **12** (2015).
19. Zhang, Z. et al. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *Journal of Proteome Research* **17**, 846-857 (2017).
20. Yu, P. et al. Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis. *Analytical chemistry* **89**, 8884-8891 (2017).
21. Deutsch, E.W. et al. Expanding the use of spectral libraries in proteomics. *Journal of Proteome Research* (2018).
22. Serang, O., Paulo, J., Steen, H. & Steen, J.A. A Non-parametric Cutout Index for Robust Evaluation of Identified Proteins. *Molecular & Cellular Proteomics* **12**, 807-812 (2013).