



OPEN

Ad hoc learning of peptide fragmentation from mass spectra enables an interpretable detection of phosphorylated and cross-linked peptides

Tom Altenburg^{1,2}, Sven H. Giese¹, Shengbo Wang^{1,3}, Thilo Muth⁴ and Bernhard Y. Renard¹✉

Mass spectrometry-based proteomics provides a holistic snapshot of the entire protein set of living cells on a molecular level. Currently, only a few deep learning approaches exist that involve peptide fragmentation spectra, which represent partial sequence information of proteins. Commonly, these approaches lack the ability to characterize less studied or even unknown patterns in spectra because of their use of explicit domain knowledge. Here, to elevate unrestricted learning from spectra, we introduce 'ad hoc learning of fragmentation' (AHLF), a deep learning model that is end-to-end trained on 19.2 million spectra from several phosphoproteomic datasets. AHLF is interpretable, and we show that peak-level feature importance values and pairwise interactions between peaks are in line with corresponding peptide fragments. We demonstrate our approach by detecting post-translational modifications, specifically protein phosphorylation based on only the fragmentation spectrum without a database search. AHLF increases the area under the receiver operating characteristic curve (AUC) by an average of 9.4% on recent phosphoproteomic data compared with the current state of the art on this task. Furthermore, use of AHLF in rescoring search results increases the number of phosphopeptide identifications by a margin of up to 15.1% at a constant false discovery rate. To show the broad applicability of AHLF, we use transfer learning to also detect cross-linked peptides, as used in protein structure analysis, with an AUC of up to 94%.

Publicly available mass spectrometry (MS)-based proteomics data have grown exponentially in terms of the number of datasets and amount of data¹. This is because high-throughput proteomic studies generate a vast pool of fragmentation spectra. Each spectrum contains characteristic peak patterns that appear due to the fragmentation of a given peptide. These peptides are used to study the proteins contained in the biological sample. It might seem obvious to apply deep learning to solve various problems with this wealth of data, but the direct application of deep learning to fragmentation spectra has not yet sparked in the community. Instead, the MS wet lab workflow is usually followed by a conventional database search². In a search, each acquired mass spectrum is scored against a list of candidate peptides from *in silico* digested proteins. For each peptide candidate, a theoretical spectrum is constructed and compared to the acquired spectrum³.

The identification of spectra remains challenging, as proteins are often either mutated or carry post-translational modifications (PTMs). The latter are essential for various biological processes, and protein phosphorylation is an important PTM that regulates protein function and facilitates cellular signalling^{4,5}. Various sophisticated algorithms exist to cope with the challenges that arise from PTMs and mutations, but these algorithms still require protein databases^{6–9}. Some attempted predictions, for example to detect a PTM, are based on only the spectrum itself and are therefore independent of a database. However, current approaches are based on engineered features and classical machine learning^{10,11}. A fragmentation spectrum can contain PTM-specific patterns (for example relations between peaks in a spectrum) that coexist with fragments resulting from the plain peptide sequence¹². These patterns can even appear in an equivariant manner, that is, they can pinpoint the

position of a PTM within the peptide sequence, but their presence alone can reveal the PTM itself¹³. Most importantly, the detection of a PTM can be separated from the sequence retrieval, and a deep learning approach would account for the variety and complexity of PTM-specific patterns.

Aside from this, there is a plethora of open challenges in MS-based proteomics, including the detection of PTMs¹², prediction of phosphosite localization scores¹³, detection of cross-linked peptides¹⁴, characterization of the dark matter in proteomics by assigning spectrum-identifiability scores¹⁵, augmentation of features for post-search rescoring¹⁶, identification of biomarkers¹⁷ and detection of anomalies including non-proteinogenic amino acids¹⁸, to name a few. These may be solved by a deep learning approach when the underlying model is able to gather biochemically relevant reasoning from a large pool of spectra. As a proof of concept, we tackle three of the challenges mentioned above, namely the detection of phosphorylated peptides based on their spectra (AHLFp), the detection of cross-linked peptides based on their spectra (AHLFx) and an improved rescoring of peptide spectrum matches.

Regarding biological data, deep learning approaches on imaging or sequential data are very successfully published at a high frequency. We argue that this is mainly due to the straightforward applicability of findings from computer vision and natural language processing to medical image¹⁹ or genomic data²⁰. Interestingly, there are at least two applications of deep learning models that are well received by the proteomics community, namely fragment intensity prediction and *de novo* sequencing. Yet, the models that are used for both applications are built around the concept of having a peptide sequence (as input, output or intermediate representation) rather than the spectrum alone. A recurrent neural network applied to a

¹Hasso Plattner Institute for Digital Engineering, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany. ²Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. ³European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK. ⁴Federal Institute for Materials Research and Testing (BAM), Berlin, Germany. ✉e-mail: bernhard.renard@hpi.de

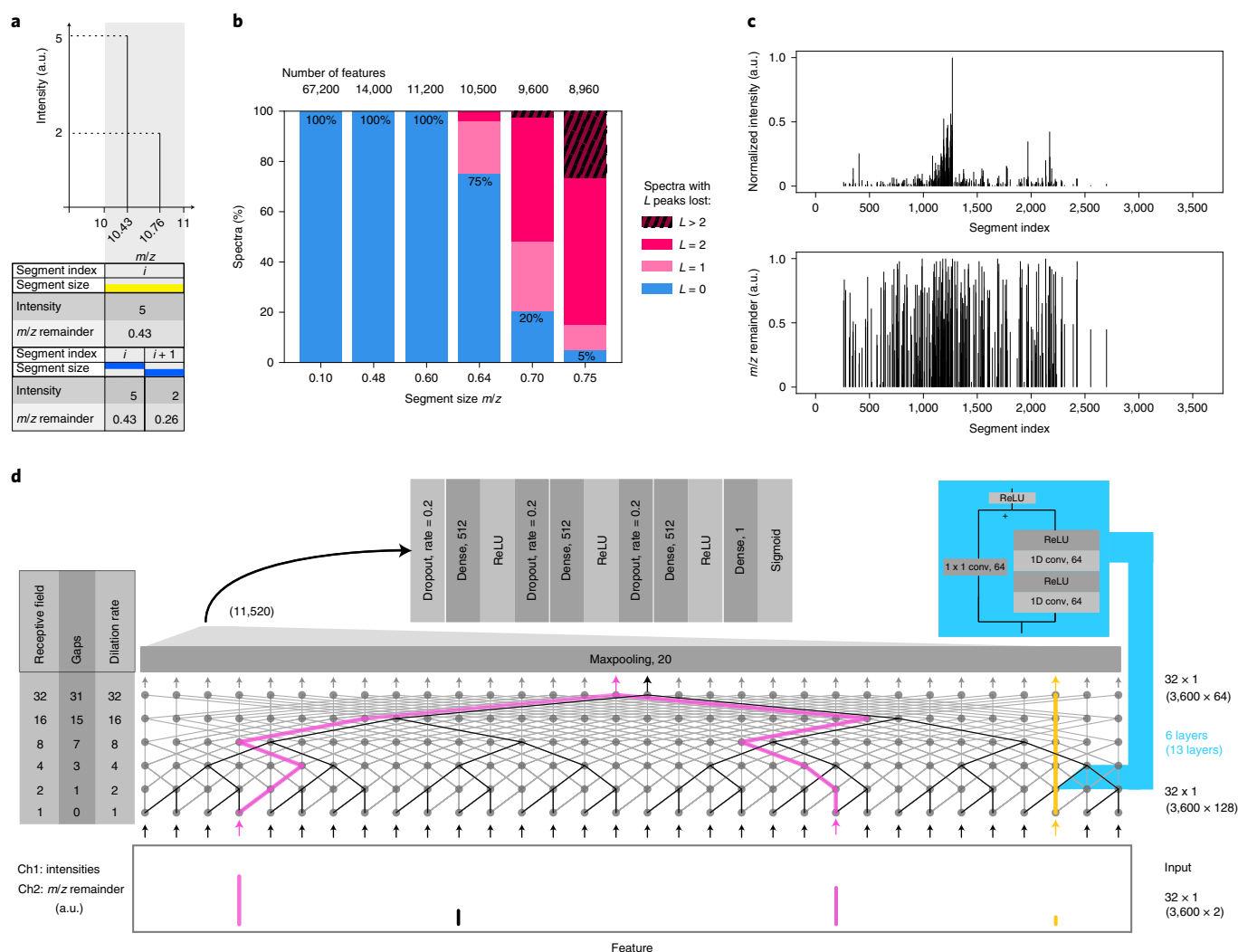


Fig. 1 | Two-vector spectrum representation is memory-efficient but keeps exact m/z locations. **a**, Top: example spectrum of two hypothetical peaks. Bottom: feature representation as two-vector spectrum, for a too large and thus lossy ($1m/z$) segment size (yellow) and for a smaller, loss-less ($0.5m/z$) segment size (blue). **b**, Trade-off between number of features (top, with a window size of 100–3,560 m/z) and loss of peaks (y axis), depending on the chosen segment size (x axis). Note that the striped area at the top ($L > 2$) of the stacked bars for the two widest segment sizes reflects spectra with more than two peaks lost. **c**, Fragmentation mass spectrum represented in two vectors, namely intensity (top) and m/z remainder (bottom). The two feature vectors have 7,200 features in total. From the m/z remainder the original m/z values can be fully recovered. **d**, Illustration of how long-range associations can be learned by AHLF via dilated convolutions. The receptive field grows exponentially with additional layers according to specific dilation rate (left box). Learnable associations are highlighted as coloured paths (grey, black, pink and yellow) (see Methods for details). Note that only selected parts of the actual network are illustrated here. In parentheses on the right, the actual tensor size and numbers of filters are compared against what is illustrated here (without parentheses). One channel of a hypothetical two-vector representation (bottom) with four peaks is shown here, whereas two actual channels (Ch1 and Ch2) are presented to the model AHLF. 1D, one-dimensional; conv, convolutional layer; ReLU, rectified linear unit; dense, fully connected layer. Hyperparameters of AHLF are summarized in Supplementary Table 1.

peptide sequence predicts fragment intensities as implemented in pDeep²¹ and Prosit²². Similarly, a recurrent neural network facilitates de novo sequencing by guiding a dynamic programming approach with a learned heuristic of peptide sequence patterns in the case of DeepNovo^{23,24} and PointNovo²⁵. These approaches also use convolutions or t-nets, but remain unable to characterize basic fragmentation patterns on their own, that is, the various ion types of amino acids are built-in and not trainable parameters.

Connected to this is the lack of interpretability for these models because the built-in features prevent any further interpretation altogether. Predefined features may reflect the experts' experience but restrict the flexibility of the model and effectively reduce interpretability²⁶. Here AHLF is provided with the entire spectrum to gain

an unbiased understanding of how peptides fragment and consequently to improve the interpretability of AHLF at the peak level.

To the best of our knowledge, there have been no attempts to directly present fragmentation mass spectra to a deep learning model and to ad hoc learn fragmentation patterns. Here, ad hoc (Latin: 'for this specific purpose') learning means that AHLF is able to abstract fragmentation patterns from spectra that are essential, for example to detect phosphorylated peptides based on their fragmentation spectra. The notion of ad hoc learning is covered by the term 'deep learning' already. However, it emphasizes that our model is able to recognize relevant and essential patterns in spectra without ever being explicitly told about them. To the best of our knowledge, current learning-based approaches in the field of MS-based

Table 1 | Performance of models in detecting spectra of phosphopeptides

Model	ROC-AUC	F1-score	Bacc
Median			
AHLFp	88.48	83.32	76.80
PhoStar	80.01	58.38	69.25
Mean			
AHLFp	85.98	74.25	76.06
PhoStar	80.99	61.32	72.95
Variance			
AHLFp	0.85	4.88	0.93
PhoStar	2.45	8.33	2.73
Overall			
AHLFp	92.09	85.51	83.68
PhoStar	91.68	79.89	82.00

Median, mean and variance for ROC-AUC, F1-score and Bacc over the individual datasets are shown at the top. Overall ROC-AUC, F1-score and Bacc as calculated over all holdout predictions without averaging on the level of cell or tissue types are shown at the bottom. Bold indicates the best model performance.

proteomics of similar scope provide masses of amino acids, ion types, losses, or combinations thereof to their models before training, which is not necessary for AHLF. For example, a PTM detection would require the learning of modification-specific features, which we can further investigate after a model has been trained.

Therefore, we propose 'ad hoc learning of peptide fragmentation' (AHLF), an end-to-end trained deep neural network that learns from fragmentation spectra to perform versatile prediction tasks. To perform these challenging prediction tasks, AHLF features a framework enabling efficient training on large numbers of fragmentation spectra, the learning of long-range peak associations through dilated convolutions, true end-to-end training on spectra due to not including domain knowledge, and interpretation of AHLF investigating whether biochemically relevant patterns are recognized.

We evaluate our approach on previously published phosphoproteomic data, which include datasets from more than one hundred public repositories⁴. In addition, we validate the results on recently published data^{27–31}; in the case of cross-linking data, we use previously published data to evaluate our transfer learning approach^{32–34}.

We interpret our model by comparing peptide fragments from ground truth peptide identifications against feature importance values for AHLF, which we calculate for each peak per spectrum individually on a collection of spectra. This is enabled by applying the Shapley additive explanations (SHAP) framework³⁵ to predictions from AHLF. In addition, we show that AHLF recognizes biochemically relevant fragmentation patterns. In this case, we do not require annotations from identified peptides beforehand. We achieve this by computing pairwise interactions, using Path Explain³⁶, between any two peaks per spectrum, and subsequently identify relevant delta masses between respective peak pairs.

We demonstrate the broad scope of our approach by applying AHLF to a distinct task, namely the detection of cross-linked peptides (AHLFx). Cross-linking is used to study the structure of single proteins, multiprotein complexes or protein–protein interactions³⁷. Detecting spectra from cross-linked peptides is challenging and different from PTM detection, because in this case two peptides including the cross-linker molecule are present in the same spectrum and need to be detected.

Finally, we show that AHLF predictions improve the number of peptide identifications in a rescoring approach. Here, phosphoproteomic datasets are reanalysed in a localization-aware open search

using MSFragger^{6,38} and then rescored by Percolator¹⁶ using AHLFp scores. This improves the number of identified peptides at constant false discovery rate (FDR). In addition, AHLFp improves the number of identification at constant false localization rate (FLR) as estimated by LuciPHOr2^{39–41}. A similar approach is shown for AHLFx on cross-linking data improving the number of cross-linked peptide spectrum matches.

Results

AHLF promotes learning of long-range peak associations. A key challenge for deep learning on proteomics data is the representation of spectra. AHLF exploits the sparsity of fragmentation spectra to derive a memory-efficient representation that accounts for exact peak locations (Fig. 1). In particular, we propose a two-vector representation, holding intensity and mass-over-charge (m/z) remainder information (Fig. 1a,c, Supplementary Note 1 and Supplementary Algorithm 1). The original spectrum can be recovered from this representation while the number of actual features is reduced (Fig. 1b). Furthermore, the two-vector representation allows the use of deep learning models such as AHLF.

In general, peaks of peptide fragments are scattered over an entire spectrum. Hence, we design our deep learning model to promote the learning of associations between any peaks while respecting their location within a spectrum (Fig. 1d). Therefore, we use convolutions with gaps, commonly called dilated convolutions. As a result, our network has a receptive field that spans the entire feature vector. In particular, the receptive field grows exponentially with the number of layers facilitated by dilations (Fig. 1d, Supplementary Fig. 1 and Supplementary Notes 2 and 3).

AHLFp detects phosphopeptides from fragmentation spectra.

Here we exemplify our approach on a specific-use case by applying AHLF to phosphoproteomic data and detecting spectra of phosphorylated peptides (AHLFp). We compare our approach against PhoStar¹¹, which is a random forest model that includes carefully generated phospho-specific features. By contrast, AHLFp was applied to the data in a plug-and-play manner as AHLFp did not require any domain knowledge beforehand. Rather, AHLFp has to come up with domain-specific features on its own, which we further investigate by interpreting AHLFp later in this work.

To demonstrate the ability of AHLFp in detecting spectra of phosphorylated peptides, we evaluate the performance of AHLFp on 19.2 million labelled spectra from 112 individual PRIDE repositories (PXD012174 (ref. 4)) containing 101 cell or tissue types. The dataset is roughly balanced and includes 10.5 million phosphorylated and 8.7 million unphosphorylated peptide spectrum matches (PSMs). We perform a fourfold cross-validation yielding four independently trained deep learning models AHLFp- α , AHLFp- β , AHLFp- γ and AHLFp- δ with their respective holdout folds a, b, c and d (Supplementary Tables 2 and 3). The following results were computed by applying each model to its respective holdout fold. For convenience we refer to AHLFp in all cases. We compute binary prediction scores as well as balanced accuracy (Bacc), F1-score and area under the receiver operating characteristic curve (ROC-AUC) for each of the 101 individual datasets. As an aggregation over these sets we show mean, median and variance in Table 1 (detailed metrics for individual datasets are given in Supplementary Table 5). AHLFp showed a better performance on average compared to PhoStar¹¹ in the detection of spectra of phosphorylated peptides (for evaluation details, see Methods). For example, AHLFp achieved a higher median ROC-AUC than PhoStar (88.48 versus 80.01) while also showing lower variances (0.85 for AHLFp versus 2.45 for PhoStar). Similarly, AHLFp outperforms PhoStar on the F1-score and Bacc metrics. Furthermore, we investigate the robustness of performance of AHLF in comparison to PhoStar (Supplementary Note 4 and Supplementary Fig. 2).

Table 2 | Validation on recently published phosphoproteomic data

Dataset	Fragmentation/ MA/ CE	Bacc (AHLFp PhoStar)	F1-score (A P)	ROC-AUC (A P)
JPST000685	CID/ITMS/35	0.73 0.57	0.71 0.40	0.83 0.66
JPST000703	CID/ITMS/35	0.71 0.55	0.67 0.31	0.79 0.61
PXD013868	HCD/FTMS/25	0.87 0.93	0.94 0.97	0.95 0.98
PXD014865	CID/ITMS/35, HCD/FTMS/45	0.87 0.81	0.88 0.77	0.94 0.90
PXD015050	CID/ITMS/35	0.77 0.66	0.77 0.59	0.87 0.76

Performances are shown next to MS settings for fragmentation including collision-induced dissociation (CID) or higher-energy C-trap dissociation (HCD), mass analyser (MA) of an ion trap mass spectrometer (ITMS) or a Fourier-transform mass spectrometer (FTMS), and collision energy (CE). Bacc, F1-score and ROC-AUC are shown for AHLFp (A) and PhoStar (P). Bold indicates the best model performance.

We validate these findings by performing predictions on other phosphoproteomic datasets. We collected five recently published datasets containing nonhuman samples. AHLFp was trained on data from human samples only (see above). Furthermore, the validation data resembles spectra that stem from four different instrument types (Supplementary Table 4). AHLFp performs better in four out of five datasets; specifically, the ROC-AUC is 9.4% higher on average (Table 2). PhoStar reaches its best performance on a particular dataset (PXD013868) with an ROC-AUC of 0.98 whereas AHLFp achieves a comparable ROC-AUC of 0.95. AHLFp appears more robust overall as its lowest ROC-AUC reads 0.79; but ROC-AUC is down to 0.61 for PhoStar on JPST000703.

AHLFp interpretations distinguish fragment ions from noise.

Here we investigate whether AHLFp is basing its decision on parts of the spectrum that belong to actual fragment ions (peptide-related peaks) or rather on peaks that are considered noise ions (ions that are not explained by the underlying peptide). Therefore, we consult the original database search results, which assigned a peptide to each identified spectrum. We compare those ground truth fragment ions against peaks that appear important to AHLFp for each spectrum individually. Hence, we calculate peak-level feature importance values (SHAP values) for individual spectra and compare them to the matching ions from the identified peptide (according to database search results). This is shown for a specific spectrum in Fig. 2a. For a quantitative comparison, we calculate a SHAP-value ratio, which is the sum of SHAP values of matched ions divided by the sum of all SHAP values. Intensity ratios are calculated accordingly. Both types of ratios are illustrated in Fig. 2a and explained in Methods.

A visual inspection of the SHAP-value ratio versus intensity ratio for all spectra in the HEK293 dataset indicates that SHAP-value ratios seem to be overall higher than their intensity-based counterparts (Fig. 2b). This means that AHLFp can distinguish between fragment ions and noise ions. This separation of signal and noise is not equally prominent throughout different datasets; for example, in the OVAS dataset it is less obvious and is therefore shown as a counterexample in Fig. 2d,e. To investigate whether this signal-versus-noise separation appears throughout datasets, we perform a Wilcoxon signed-rank test on the 25 datasets from the first fold from the cross-validation splits (fold a and the respective model AHLFp- α ; see Supplementary Tables 2 and 3).

According to the Wilcoxon test, for 18 of the 25 datasets the SHAP-value ratios are statistically higher than intensity ratios, considering a one-sided significance level of $\alpha=0.01$ (Bonferroni-corrected). This supports our observation of AHLFp learning a signal-versus-noise abstraction that is correct in most datasets. To check whether these interpretations depend on the quality of spectra annotations (Fig. 2b–e), we perform the Wilcoxon signed-rank test for six score thresholds ranging from 40 to 140 (Fig. 3a). Similarly, we check whether the comparison depends on considered ion types (Supplementary Note 5). Altogether, in the

tested situations AHLFp was able to separate fragment ions from noise ions for a vast pool of spectra in the majority of datasets.

AHLFp interactions reflect biochemically relevant patterns. To investigate whether AHLFp gained further insights (additional to the signal-to-noise distinction above) we compute pairwise interactions³⁶ in combination with delta m/z between any peaks. In particular, we check whether delta masses that are relevant in the context of phosphopeptides are recognized by AHLFp (details in Supplementary Note 6). Therefore, we collect the identified phosphopeptide spectra from an individual run of the HEK293 dataset. We keep spectra that AHLFp predicted as being phosphorylated, and calculate pairwise interactions and respective delta m/z for each spectrum (Methods).

At first glance, pairwise interactions reveal a set of distinctively higher interaction strengths at certain delta m/z values (Fig. 3c,d). For example, a prominent interaction at 98 m/z reflects phosphoric acid (interaction annotated with ‘*’ in Fig. 3d). To identify the other interactions, we search delta m/z including combinations of serine, threonine or tyrosine (S/T/Y), phosphoric acid (abbreviated with ‘*’), and of loss of ammonia (NH_3) and/or loss of (H_2O) subject to charges between +1 and +4 (for detailed stoichiometry, see Methods). Hence, a total of 2,352 hypothetical delta m/z were tested against the top-30 highest interactions. Of the top-30 highest interactions, 15 can be explained with specific delta masses as annotated in Fig. 3c,d. This is controlled by assessing the matching tolerance with a random baseline and a negative baseline (Fig. 3b and Supplementary Note 6).

AHLF improves database search results through rescoring.

Here we investigate whether predictions by AHLF can be used to improve the number of peptide identifications from database searches. Therefore, we use AHLFp to assign a prediction score for each fragmentation spectrum in PXD014865 and rescore search results by using Percolator (Supplementary Note 8). The use of AHLFp increases the number of identified peptides by 8.4% and phosphopeptides by 15.1% at a constant FDR of 1% (Fig. 4a and Supplementary Fig. 6).

Furthermore, we use the rescored PSMs from above and subsequently estimate an FLR by using LuciPHOr2 (Supplementary Note 8 and Supplementary Figs. 7, 10–13). At an FLR of 1%, AHLFp increases the number of identifications by 30.4%, when the LuciPHOr2 HCD model is used (Fig. 4c and Supplementary Fig. 7) and by 7.4%, when the LuciPHOr2 CID model is used. To check whether these newly gained peptide identifications are valid, we compare a representative spectrum (newly gained by using AHLFp in PXD014865) to a corresponding reference spectrum (Supplementary Fig. 9).

By using transfer learning, we can apply AHLF to other types of data. In particular, we apply AHLF to cross-linking data, which results in AHLFx (Extended Data Fig. 1, Extended Data Table 1 and Supplementary Note 7). We use AHLFx on the PXD012723

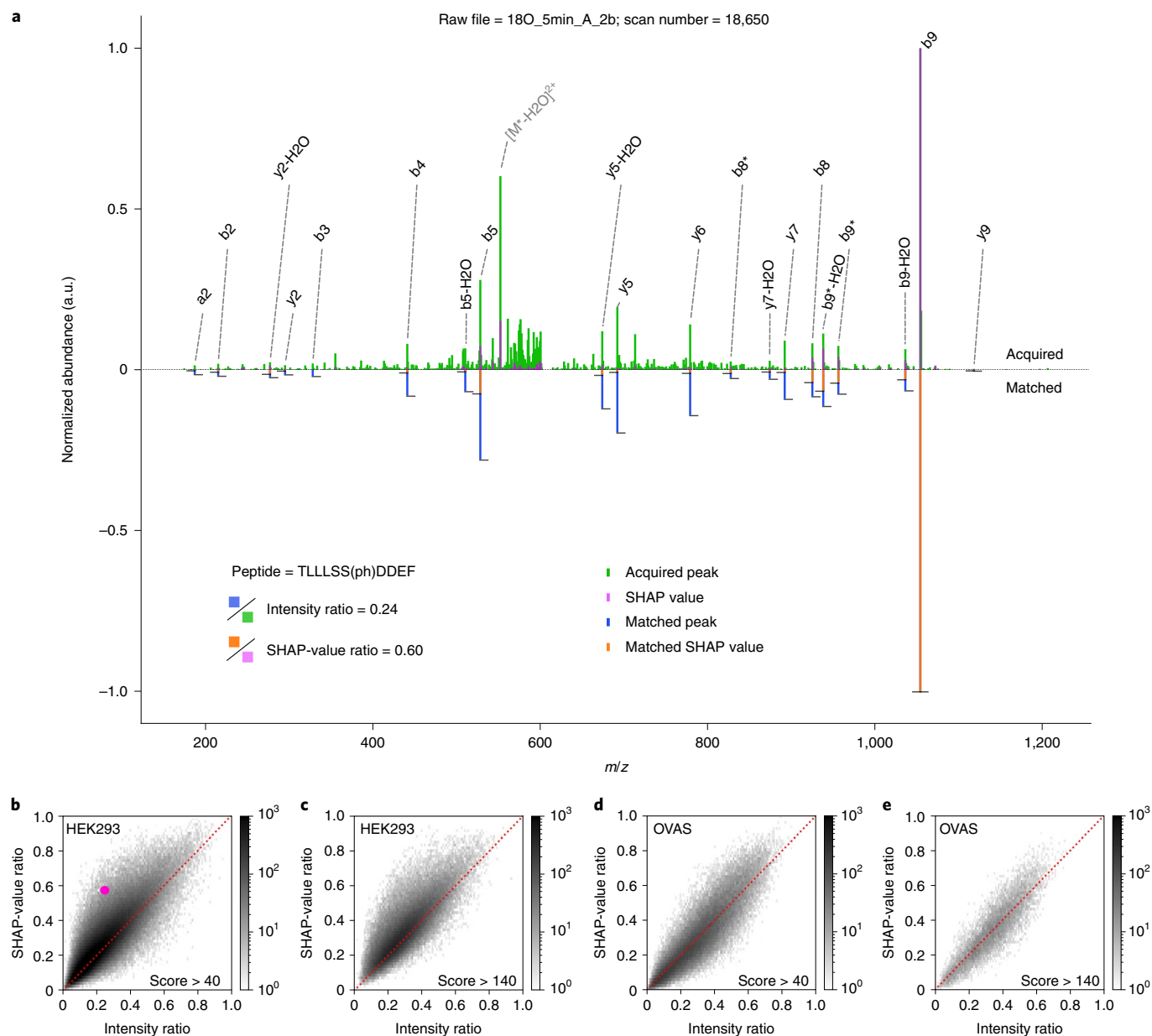


Fig. 2 | AHLFp can distinguish between fragment ions and noise ions. a, Mirror spectrum drawn from the HEK293 dataset as an example. Positive values are acquired intensities (green) and SHAP values from AHLFp (magenta). Respectively, negative values are matched ions (blue; according to the database search results) and the matched SHAP values (orange; subset of SHAP value that are at the m/z positions of matched ions). Whiskers make SHAP values (pointing left) visually distinguishable from corresponding intensities (pointing right). On the bottom left, the nominator and denominator for each of the two ratio types are indicated. Matched ions are labelled as annotated in the MaxQuant search results (we manually annotated the matching precursor ion $[M^*+H_2O]^{2+}$, which was not considered in the original search). **b–e**, Hexbin plot comparing SHAP-value ratios to intensity ratios as explained in the main text and illustrated in **a**. Each hexbin represents a collection of ratios from a set of spectra, and the number of spectra in each hexbin is given as logarithmic grey code. The pink dot in **b** (SHAP-value ratio = 0.60 and intensity ratio = 0.24) indicates the bin that contains the example spectrum shown in **a**. **c**, Same as **b** but spectra with lower scores (≤ 140) are removed. **d, e**, Same as **b** and **c** but for the OVAS dataset as a counterexample, where the SHAP-value ratios are not statistically higher than the intensity ratio (see Wilcoxon test below), which is not obvious from visual inspection alone.

dataset alongside the search results from xiSEARCH. Here we filter the spectrum matches based on the AHLFx score before applying an FDR threshold of 5% using xiFDR. This increases the number of identified cross-linked peptide spectrum matches (CSMs) by 11.2% at the optimal cut-off for the AHLFx score (Fig. 4b). Interestingly, the local maxima of the AHLFx score distribution (Fig. 4b) correspond with PSMs and CSMs (Supplementary Fig. 8).

Discussion

We present a novel method for detecting PTMs and cross-linked peptides based on their fragmentation mass spectra independent of a database search. Our approach showed a high and robust performance over a wide range of datasets. We demonstrated that AHLF has learned substantial and fundamental fragmentation-related features, which was enabled by our strict end-to-end training scheme. Interpretations of AHLF were in line with the majority of ground

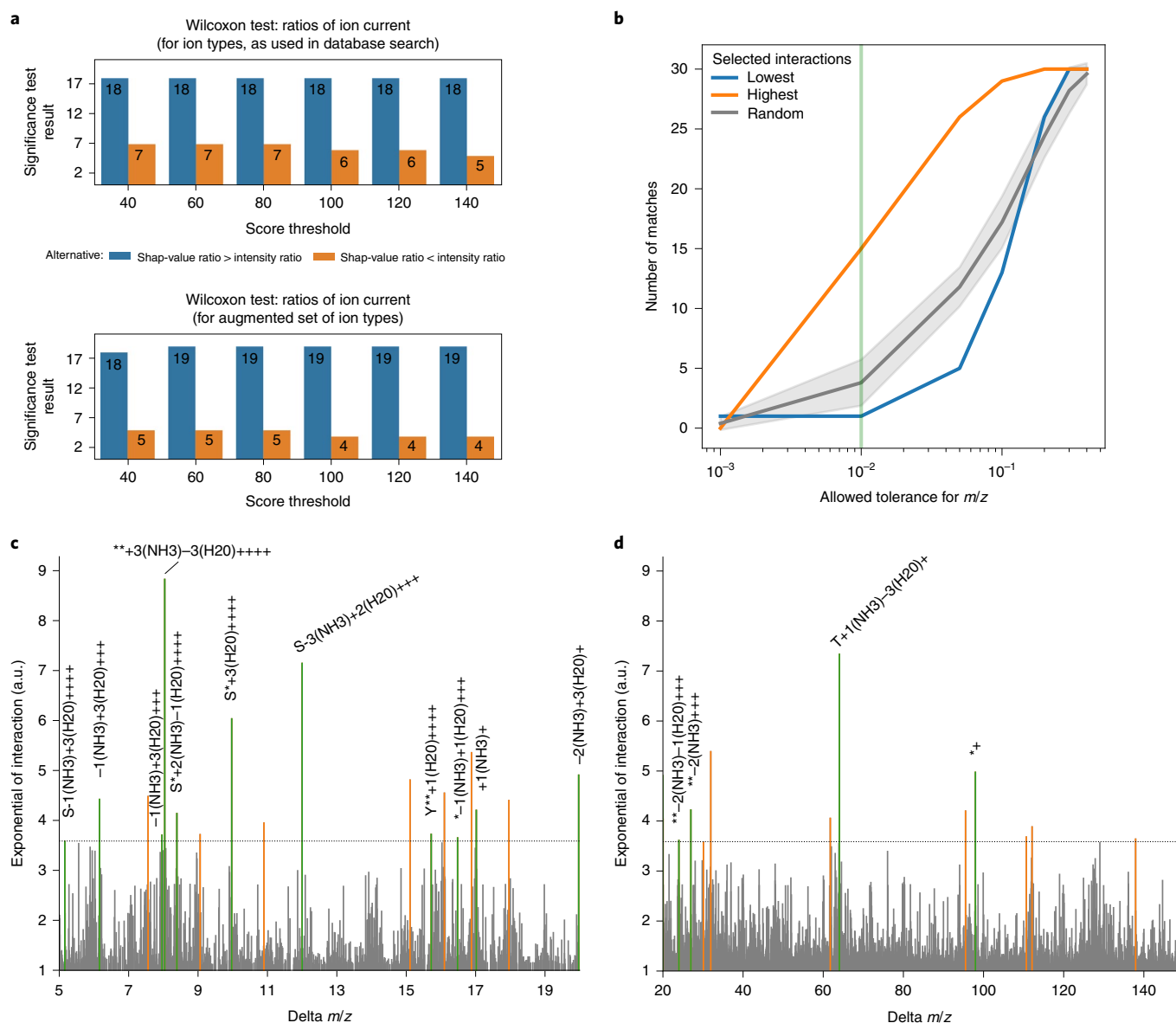


Fig. 3 | AHLF recognizes fragments beyond the classical b-/y-ions including biochemically relevant peak associations. a, Number of datasets that pass the Wilcoxon signed-rank test depending on the applied score threshold for the two alternatives, namely SHAP-value ratio > intensity ratio (blue) or SHAP-value ratio < intensity ratio (orange) for ion types, as used in the database search (top) and using the augmented set of ion types (bottom). **b**, Adjusting the maximally allowed tolerance to $0.01 m/z$ (green vertical line). Comparatively, the top-30 highest interactions (orange) and the potential negatives estimated by the bottom-30 lowest interactions (blue) and by 30 randomly sampled interactions (grey; mean and standard deviation) are shown. **c,d**, Pairwise interactions incorporate a combination of S/T/Y, phosphoric acid (*) and potential loss of ammonia (NH_3) and/or water (H_2O). Top-30 highest interactions (above dashed line) of which 15 meet the required tolerance (green) are explained as annotated. The remaining 15 interactions (orange) do not meet the required tolerance even though they still reflect reasonable interactions (Supplementary Fig. 4). (Panels **c** and **d** show the same data but were separated into two panels for a better readability and to zoom into delta m/z between $5 m/z$ and $20 m/z$. Only the highest 10,000 interactions are shown here).

truth peptide identifications. Furthermore, we could use AHLF for rescoring and thus improve the number of peptide identifications. We demonstrated the flexibility of AHLF by applying it to cross-linking data, and performed this detection task based on spectra alone.

For our approach we developed a specific two-vector representation of spectra. This allowed us to use convolutional layers and to reduce memory consumption without obscuring resolution-related information (Fig. 1; see Supplementary Note 9 for further discussion).

To model fragmentation patterns, we set up a deep neural network that we designed to promote learning of long-range

associations between features (Fig. 1d). Most biochemically related peak associations are long-range relations because peaks in a spectrum can be a hundred Dalton (Da; atomic mass for a singly charged ion) or even multiple hundreds of Da apart (for example several phosphosites at different locations of a peptide). Similarly, in our two-vector spectrum representation related features can be multiple hundreds of steps apart (Figs. 1 and 2). Hence, AHLF falls into the category of temporal convolutional neural networks^{42,43} (discussed further in Supplementary Note 10).

AHLF outperforms the current state-of-the-art PhoStar¹¹ on phosphopeptide detection on recently published datasets from diverse lab environments and experimental set-ups. In particular,

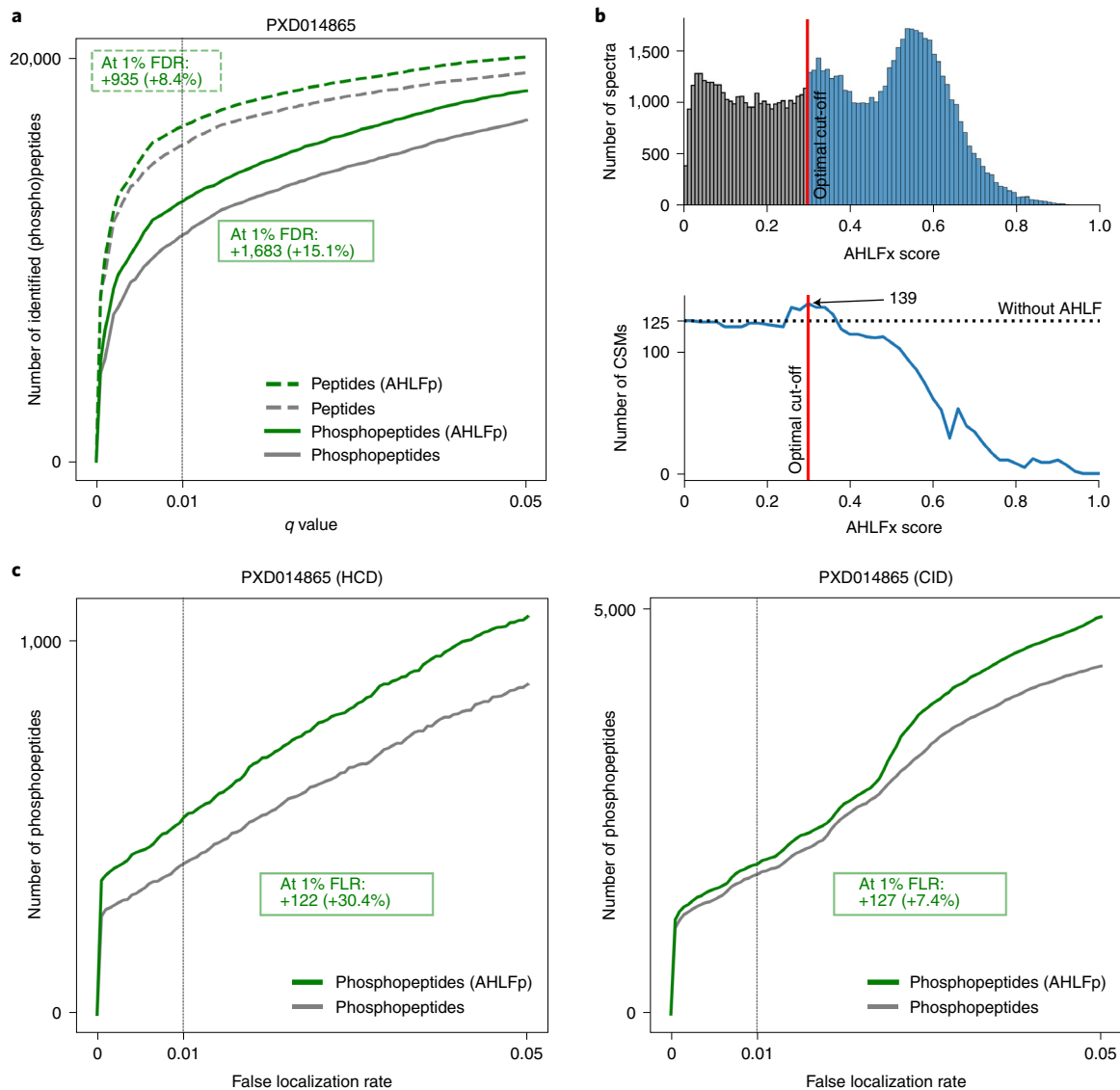


Fig. 4 | AHLFp and AHLFx improve the number of identifications when used in rescoring search results. a, Rescored MSFragger results as derived by Percolator when AHLFp scores are used (green lines). In comparison, MSFragger search results are rescored without using AHLFp (grey lines). The increase in identifications at a constant q value (FDR) of 1% (vertical dotted line) is shown in terms of absolute and relative numbers (green boxes). **b**, xiSEARCH spectrum matches versus cut-offs based on AHLFx scores before FDR estimation. At the optimal cut-off (red vertical line) the number of CSMs at 5% FDR (CSM-level) is maximized. **c**, Number of identifications versus FLR as estimated by LuciPHOR2 based on the rescored identifications of phosphopeptides using AHLFp (green lines). In comparison, an FLR is estimated for rescored phosphopeptides without using AHLFp (grey lines). Identifications are shown according to the used fragmentation method as LuciPHOR2 provides dedicated models for HCD (left) and CID (right). The increase in identifications at a constant FLR of 1% (vertical dotted line) is shown in terms of absolute and relative numbers (green boxes).

we could demonstrate that AHLF is more robust against variables such as sample species, type of instrument and dataset size. The latter is an indirect measure of lab diversity, which is mostly covered by smaller datasets where our model was showing a robust performance (Table 1). We investigated the influence of the used fragmentation type, mass analyser and collision energy on the performance of AHLF per MS run in the PXD012174 dataset (Supplementary Fig. 3 and Supplementary Note 11).

Interpretation of AHLF and subsequently investigation of the mechanisms behind what AHLF has learned coincided with biochemically reasonable fragmentation patterns. We investigated, after AHLF was trained, whether AHLF is picking up reasonable features from a given spectrum. Therefore, we derived feature importance values on the level of individual spectra for a large collection of

spectra. We could show that AHLF was using the entire spectrum (instead of picking up single peaks) to derive its predictions score, as prominent SHAP values are often scattered over the entire m/z range (Fig. 2a). Furthermore, important features coincide with peptide fragment ions. This was true for most spectra from the tested datasets (Fig. 3a and Supplementary Note 12). We could quantify that AHLF was actually learning a suitable abstraction of a given spectrum and thus was able to focus on fragment ions despite not knowing the peptide sequence in advance.

Following up on this, we computed pairwise interactions between any pair of peaks in a spectrum. These interactions point out that AHLFp has a fundamental understanding of how phosphopeptides fragment, and respective delta masses coincide with commonly known neutral losses such as phosphoric acid and combinations

thereof (Fig. 3c,d). This is analogous to engineered features that are used in Colander¹⁰ and Phostar¹¹. By contrast, AHLFp is ad hoc learning them from the data, and here we checked whether, after training, those coincide with common expert knowledge. For example, the fragmentation of a phosphopeptide often involves the loss of phosphoric acid H_3PO_4 , which results in a delta mass of around 98 Da between pairs of ions with a single positive charge^{13,44} and is indicated by ‘*’ in Fig. 3d.

In addition, we could explain 15 of the top 30 highest interactions when matching delta masses with an allowed tolerance of only 0.01 m/z (Fig. 3b) and up to 26 when choosing a slightly higher tolerance of 0.05 m/z (Supplementary Fig. 4). Conversely, the delta mass of phosphoric acid undergoing an additional loss of water (a common phospho-specific loss at around 80 m/z) was not among the selected top 30. However, a substantial interaction is recognizable at 80 m/z (Fig. 3d). Furthermore, we observe reasonable but rather complex combinations of losses for multiple charged ions^{45,46}. As these should be less common, we double-checked that by performing the same kind of analysis but for data that had been acquired on an Orbitrap instrument involving a higher-resolution mass analyser (Supplementary Fig. 5 and Supplementary Note 13).

We demonstrated a broader applicability of our framework by using transfer learning on cross-linking data⁴⁷ (Supplementary Note 14). This is crucial for situations in which training data are limited. Similarly, we performed a fine-tuning to further boost performance, for example for a specific instrument. In particular, we fine-tuned AHLFp on Q Exactive data (Extended Data Table 2).

Furthermore, we could show that AHLFp improves the number of identifications when rescoring the search results. Therefore, we did a reanalysis of the phosphoproteomic validation datasets using MSFragger and performed a rescoring by integrating AHLFp predictions (see Supplementary Note 15 for further discussion). Similarly, we could use AHLFp predictions to filter spectra before FDR estimation by xiFDR. At the optimal cut-off of AHLFp scores we could maximize the number of identified CSMS at 5% FDR (Fig. 4b). Overall, these experiments show that AHLFp predictions are largely orthogonal to search results alone, and predictions derived from raw spectra using our deep learning model AHLF complement and improve existing MS workflows. Based on these rescoring results, we anticipate that the idea of rescoring could be further extended by using AHLF in combination with its SHAP values (Fig. 2a). The rescoring could be augmented by peak-level features because the SHAP values by AHLF would pinpoint parts per individual spectrum that are relevant for inclusion in rescoring.

We foresee that our approach will spark diverse future applications as feature creation is fully integrated in the learning process. Possible future applications include the prediction of phosphosite localization scores, spectrum identifiability scores, further augmentation of post-search rescoring, biomarker detection, and anomaly detection including the detection of non-proteinogenic amino acids and uncommon PTMs. This deep learning model is one of the first to be able to ad hoc learn the fragmentation patterns in high-resolution spectra.

Methods

Public datasets used for cross-validation. For training and testing through cross-validation, we used the combined dataset PXD012174 (ref. 4), which contains 112 individual repositories organized according to 101 human cell or tissue types (number of zip files in PXD012174 that contain fragmentation spectrum raw files). The individual datasets used phospho-enrichment assays. This combined dataset was reanalysed by Ochoa et al.⁴ and underwent a joint database search using MaxQuant^{48,49} with the following error rates: FDR set to 0.01 at PSM, protein and site decoy fraction (PTM site FDR) levels. The minimum score for modified peptides was 40, and the minimum delta score for modified peptides was 6. The combined search results were taken from the ‘txt-100PTM’ search results, as described in PXD012174 (ref. 4). For each spectrum, a label (phosphorylated or unphosphorylated) was assigned when the set of PSMs contained exclusively

phosphorylated or unphosphorylated peptides. In other words, if the set of identifications for a given spectrum contained PSMs for both kinds of peptides (phosphorylated and unphosphorylated), then we discarded the spectrum. By using this strategy, we reduced labelling errors. Overall, this yielded a training set that contained 19.2 million PSMs consisting of 10.5 million phosphorylated PSMs and 8.7 million unphosphorylated PSMs.

Model training. We optimized the cross-entropy loss using the adaptive learning rate optimizer ADAM⁵⁰ with an initial learning rate of 0.5×10^{-6} . AHLFp was trained for 100 virtual epochs consisting of 9,000 steps, which was a gradient descent step on a mini-batch of size 64. For regularization we used early stopping and dropout with a rate of 0.2 on the fully connected layers. We initialized weights for all layers using ReLU activation with random weights drawn from a standard normal distribution using He correction, and we used Glorot correction for layers not using ReLU activation (Supplementary Note 3).

Public datasets used for validation. For validation we used five datasets^{27–31}. For each dataset, we used the original search results from MaxQuant alongside the raw spectra. In addition, we removed PSMs with scores lower than 40. The detailed information is summarized in Supplementary Table 4 for each dataset.

Evaluation of AHLF. We compared AHLFp against PhoStar¹¹, which is currently the state-of-the-art method for phosphopeptide prediction based on fragmentation spectra. We used PhoStar with default parameters: the m/z tolerance was set to 10 ppm, the peak-picking depth was set to 10 (per 100 m/z) and the score threshold was set to 0.5 for all results shown here. PhoStar is closed-source and not trainable by the user. We used the original PhoStar ensemble model parameters¹¹.

For 462,464 spectra from PXD012174 (2.4% of the dataset) PhoStar was technically not able to predict a classification score (it provides an error about mismatching masses). This also applied to some spectra from the validation data. We proceeded by assigning a PhoStar prediction score of 0.5 to these spectra to achieve a fair comparison.

To calculate metrics such as balanced accuracy, F1-score and the ROC-AUC, we used Scikit-learn⁵¹. We set a score threshold to 0.5 to get class labels from the predicted continuous binary classification scores. Spearman correlation coefficients were calculated by using Scipy⁵².

In the case of fourfold cross-validation the four resulting models of AHLFp were evaluated on their respective holdout dataset (Supplementary Table 3). For the benchmark on unseen, recently published data, we used the arithmetic mean of prediction scores from the AHLFp model ensemble.

Methods used for explaining AHLF. For the calculation of feature importance values we used SHAP⁵³. From the SHAP framework, we used DeepExplainer and we set an all-zeros vector as background reference spectrum. In particular, we chose SHAP because it allowed us to investigate each spectrum individually (in contrast to global methods, which report only aggregated statistics over multiple data points). Furthermore, SHAP computes importance values that are additive, which means for a given spectrum their sum is supposed to mirror the prediction score of AHLF for that spectrum. In our case, the errors between the sum of SHAP values and AHLF scores were smaller than 1%. Here we refer to the additive feature importance values as SHAP values.

Furthermore, we were interested in absolute SHAP values as we investigated both types of spectra (from either a phosphorylated or an unphosphorylated peptide) equally. In particular, we were testing whether AHLF can separate fragment peaks from noise peaks. Therefore, we assumed that for each spectrum the sum of all intensities $\sum I := \sum I_{\text{matching}} + \sum I_{\text{non-matching}}$ consists of intensities that are at an m/z that could be matched to a peptide fragment ion and consists of other intensities that could not be matched. Furthermore, we defined an intensity ratio $:= \sum I_{\text{matching}} / \sum I$ and analogous SHAP-value ratio $:= \sum |s|_{\text{matching}} / \sum |s|$. Note that these ratios are bound between zero and one and that any scaling factor cancels out such that we could compare the two types of ratios easily. We anticipated that our explainability assessment depends on which ions are matching; hence, we investigated crucial parameters that potentially alter the ground truth. In particular, we systematically looked at the quality of PSMs and also looked at the considered ion types as described below.

To have a ground truth on the level of each individual peak within a given fragmentation spectrum, we used the PSM information from the search results. For our tests using the original set of ions (as matched during the database search), we used the MaxQuant output from the MS/MS table (msms.txt file) directly. For each PSM, a set of matching peaks including m/z , intensity and ion type has been reported. The ion types that were used during the database search contained a, b, y, b-H3PO4, y-H3PO4, b-NH3, y-NH3, b-H2O and y-H2O ions. By contrast, for our experiments with an augmented set of ion types, we computed in addition the theoretical spectrum for each peptide for a given PSM. Therefore, we augmented the ion types listed above by including types a-H2O, a-NH3, c, c-dot, c-H2O, c-NH3, M, M-H2O, M-NH3, x, x-H2O, x-NH3, z, z-dot and z-H2O. We made sure to reproduce the fragment masses that were matched during the original search and concatenated the additional calculated fragment ions, yielding an augmented theoretical spectrum. To find matching peaks between an acquired and the

augmented theoretical spectrum, we used a binary search as implemented by ref.⁵³. For matching peaks we allowed a mass tolerance of either 0.5 Da in the case of an ion trap or 20 ppm in the case of an Orbitrap.

For each spectrum we were able to compute two types of ratios as stated above (both reflecting a measure of signal versus noise). We assumed that these resemble two random variables that are measured on the same spectrum. Therefore, we chose the Wilcoxon signed-rank test to compare whether one ratio is statistically greater than the other ratio. We chose a significance level of $\alpha = 0.01$ for the one-sided signed-rank test and used Bonferroni correction to account for the number of thresholds times the number of datasets as the total number of hypotheses. Test statistic and P values were computed using the Wilcoxon test from Scipy⁵².

For pairwise interactions we used Path Explain⁵⁶. Path Explain computes interactions between any pair of input features for a deep learning model, in our case any pair of peaks within a spectrum. Computationally, this is very expensive (more than one hour per spectrum). We excluded spectra with more than 500 peaks (Path Explain computations increase quadratically with number of peaks). We selected a single run, '18O_5min_A_2b' from the HEK293 dataset. Furthermore, we selected spectra from identified phosphorylated peptides (by MaxQuant) and required AHLF to correctly predict them; any other spectrum was discarded. This yielded 193 spectra for which we computed pairwise interactions. In the case of the Orbitrap run, we used '5_min_M_a_QE.raw' from the HeLa dataset. Again we kept spectra for identified phosphorylated peptides only, resulting in 504 spectra for which pairwise interactions are calculated. In any case, we kept positive interactions as we were interested in the positive class and the responsible pairwise interactions. In the case of the Orbitrap run, we used the Python package 'ms-deisotope' to de-charge and de-isotope the spectra. For each isotopic envelope (groups of peaks are summarized as one mono-isotopic peak) we averaged over the pairwise interactions by using the arithmetic mean.

To identify the interactions and assigned neutral losses, we searched delta m/z by including any combination of $[0, 1]$ of S/T/Y, $[0, +1, +2]$ of phosphoric acid (abbreviated with '*'), or $[-2, -1, 0, +1, +2]$ losses of ammonia (NH₃) and/or $[-2, -1, 0, +1, +2]$ losses of (H₂O) subject to charges between +1 and +4. These combinations resemble 2,352 different delta m/z . In the case of the Orbitrap data, the number of hypotheses was reduced to 588, because there was no need to account for different charge states. In the main text we refer to a peak-matching equivalent tolerance of 0.01 m/z . As delta masses are matched, we multiplied this tolerance by $\sqrt{2}$; that is, according to Gaussian error propagation for a difference of $mz_1 - mz_2$, the effective error is $\epsilon = \sqrt{\epsilon_1^2 + \epsilon_2^2}$, such that we used $0.01 \times \sqrt{2}$ as apparent tolerance when matching delta m/z .

Transfer learning on public cross-link data. For training and testing we used public data from JPST000916 (DSS as cross-linker) and for evaluation we used JPST000845 (BS3 as cross-linker) as holdout set^{32,33}. Raw files were converted to MGF (Mascot generic format) and after m/z recalibration were searched with xSEARCH⁵⁴. As labels we used CSMs from xiSEARCH results as positive class (at 5% CSM FDR) and the reported linear PSMs as negative class (5% PSM FDR). During transfer learning we took our pretrained models AHLFp- α -AHLFp- δ and continued training on JPST000916; we adjusted the learning rate to 0.0001 and increased the dropout rate accompanying the dense layers to 0.5. All other hyperparameters were kept the same as for the original training of AHLF. For the baseline models we used a fully connected network with 2, 3, 4, 5 or 6 layers, with 32, 288, 544, 800 or 1024 units, and with a dropout rate of 0.0, 0.2 or 0.5 using ADAM with initial learning rates of 1.0×10^{-3} , 0.5×10^{-3} , 1.0×10^{-4} , 0.5×10^{-4} or 1.0×10^{-5} . We randomly sampled 50 parameter constellations and trained each three times. From the resulting 150 training runs we evaluated the best four runs. This was repeated for the feature vector ($m/z', I$) and the two masked-out versions ($_, I$) and ($m/z', _$), where respective values were set to zero. During transfer learning and baseline training we used early stopping on the test-set split.

Spectrum representation of AHLF. With the help of our particular spectrum representation we exploited the sparsity of centroided spectra. A centroided fragmentation spectrum is a list of peaks that are tuples of m/z and intensity I . If the sparsity matches the segment size (meaning exactly one peak per segment) this operation is reversible as illustrated in Fig. 1a. In other words, the original peaks list can be recovered from our two-vector representation. To achieve a truly loss-less conversion we could exploit the sparsity of centroided spectra and adjust the segment size accordingly. Theoretically, the chance of two peaks randomly falling into the same segment is marginal. A peaks list contains l total number of peaks. Typically, an MS/MS spectrum has hundreds, or in extreme cases up to a few thousands, of centroided peaks l and a dense vector matching the instrument resolution has easily multiple hundreds of thousands of entries b . Assuming uniform sampling of values for m/z , the probability of randomly choosing an m/z (within the resolution) that has been occupied already is given by $1/(b-l)$. It is known that m/z displays dead spots (for example, combined histogram of m/z from all peaks lists). Hence, m/z is usually not uniformly distributed. However, the opposite case of a fragmentation spectrum in which all peaks are clumped together is usually discarded during quality assessment. These poorly fragmented spectra are not very informative.

In Results, we describe a strategy of how to generate our proposed spectrum representation from a given peaks list (Fig. 1a). An alternative but equivalent strategy is to first populate an all-zero vector (of size window-range times instrument-resolution) with peaks, and then apply the maximum and the argument maximum within small mass segments of fixed size (Fig. 1a). Furthermore, our spectrum representation reflects a regular grid, which is equally spaced and with fixed connectivity of m/z values. This makes a spectrum compatible with a network that uses convolutional or recurrent layers (comparable to applications including an image or a time series with constant time steps).

To handle the amount of spectra, especially for feeding a graphics processing unit with training samples, we set up a custom pipeline facilitating the conversion of peaks lists into the two-vector representation. It also performs common preprocessing steps⁵⁵, for example ion current normalization that divides intensities by the sum of squared intensities. Our preprocessing pipeline is implemented in Python, largely integrating Pyteomics^{56,57} and TensorFlow⁵⁸. The combined preprocessing and training pipeline can be found as part of our code repository. Our pipeline accepts MGF files as input files. Spectra from raw files were centroided and converted to MGF files by using ThermoRawFileParser⁵⁹.

Details about particular choices for the model architecture of AHLF. In Results, we illustrate the model architecture of AHLF (Fig. 1d). We show how a single output (Fig. 1d, black arrow at centre) can receive information from any input (black arrows) via a collection of paths (black solid lines) in the block of dilated convolutions. The block of dilated convolutions facilitates a receptive field that spans the entire feature vector. The block has nested stacks of convolutional layers and 64 filters per layer; hence, the actual model complexity is not fully captured by the simplified illustration here (blue inlay). On the right in Fig. 1d we compare what is drawn versus what was implemented (in parentheses). In addition, we illustrate how a single output can learn to reflect a specific pair of ions (pink path). Associations between more than two input features can be learned by the model, but are not illustrated here. Furthermore, we introduced skip connections by using convolutions with kernel size of 1 that are added to the stacked convolutions (blue inlay). This allows for inputs to be passed to the output (yellow path).

A block of dilated convolutions is commonly called a temporal convolutional neural network (TCN)⁶². In a TCN, the receptive field grows exponentially, and therefore the gradient computation only needs $\log(\text{distance between features})$ steps. We preferred this over, for example, a transformer architecture⁶⁰, even though the latter facilitates gradient computation that is independent of the distance between two features. However, it requires keeping a self-attention matrix, which in turn scales quadratically with input size. By contrast, for a TCN the computational complexity of convolutions scales linearly with the input size⁶⁰. Hence, in the case of AHLF we chose a TCN over a transformer.

In our TCN, we used padding that conserves the size between the input and the output layer ('same'-padding). By contrast, TCNs are also usually used in conjunction with 'causal'-padding, for example the 'Wavenet' uses causal-padding⁶¹. This is because the latter is modelling audio signals over time, and feedback from the future to the past is eliminated by causal-padding on purpose. In the case of a fragmentation spectrum, the peptide fragmentation happens from both ends of a peptide and a spectrum is bi-directional in that sense; hence, we decided to stay with same-padding. An output feature is able to receive information from the left and the right part of the previous layer (Fig. 1d). The TCN block is followed by fully connected layers. The final prediction is facilitated by a sigmoid as activation function, which outputs a score between zero (unphosphorylated) and one (phosphorylated).

MSFragger searches for the reanalysis of validation datasets. For each run from the validation datasets (JPST000685, JPST000703, PXD013868, PXD014865 and PXD015050) we assigned a prediction score to each MS/MS spectrum using AHLFp- α -AHLFp- δ . Note that PXD013868 is an extraordinarily large dataset, which contains 30 tissue types of *Arabidopsis thaliana*. Hence, for this reanalysis we performed searches for runs from the seed tissue using enrichment by immobilized metal affinity chromatography (as representative for the other tissues in PXD013868). We used MSFragger (version 3.3) and searched the spectra from JPST000685 against UP000244005_3197.fasta, JPST000703 against UP000006548_3702.fasta, PXD013868 against UP000006548_3702.fasta, PXD014865 against UP00000589_10090.fasta and PXD015050 against UP00000589_10090.fasta. Finally, for true FDR calculation, we searched spectra of PXD014865 against the target species (*Mus musculus*, UP00000589_10090.fasta) concatenated with the trap proteome (*Arabidopsis thaliana*, UP000006548_3702.fasta). These proteomes were downloaded from UniProt (<https://www.uniprot.org/>) in August 2021. For the searches, the minimum peptide length was set to 7 and maximum peptide length was set to 50, and up to two miscleavages were allowed. Oxidation of methionine, protein N-terminal acetylation, and phosphorylation of serine, threonine and tyrosine were set as variable modifications. Up to three variable modifications per peptide were allowed. Cysteine carbamidomethylation was set as a fixed modification. Fragment mass tolerance was set to 20 ppm (for FTMS runs) or 0.5 Da (for ITMS runs).

For the synthetic phosphopeptide library PXD000138 (ref.⁶²) we adapted the search parameters from ref.⁶². The HCD runs from the library were

searched against the International Protein Index (IPI) human proteome and the concatenated synthetic peptide library (as provided by ref.⁶²), allowing tryptic peptides of length between 9 and 27 with up to four miscleavages and with unmodified cysteine, up to three oxidations (methionine) and up to one phosphorylation (serine, threonine or tyrosine) per peptide.

All searches were open searches with a precursor window of size (−150 Da, +500 Da), and the `localize_delta_mass` parameter was enabled. We allowed MSFragger to report the top-10 scoring PSMs per spectrum. Results by MSFragger are exported as Percolator input (PIN) files and forwarded to Percolator for postprocessing (see below).

Rescoring and FDR estimation for the reanalysis of validation datasets. From MSFragger we exported PIN files. These PIN files contain features as specified by MSFragger. We added to the PIN files a feature column containing a Boolean that reflects whether the candidate peptide contains a phosphorylation or not. In addition, we included four feature columns according to predictions by the model ensemble AHLFP- α -AHLFP- δ . The original MSFragger features in combination with the Boolean feature are referred to as PIN files without using AHLFP (baseline), whereas all aforementioned features together are considered as PIN files with using AHLFP. We ran Percolator (version 3.05.0) using unity-length normalization with a target FDR of 1%.

FLR estimation for the reanalysis of validation datasets. For false localization rate estimation we used the Percolator output target PSMs (with or without AHLFP predictions) from above. These were filtered by a PSM-level FDR of 1% and then used as input for LuciPHOR2 (version 2.1). According to each dataset, we used either the CID model or the HCD model of LuciPHOR2. Furthermore, the MS/MS tolerance was set to 20 ppm (for FTMS runs) or 0.5 Da (for ITMS runs).

Quantification of the true FDR and true FLR. For the calculation of the true FDR we counted peptides from the trap species (in the case of PXD014865) as false positives. In the case of PXD000138, peptides from the IPI human proteome (but not included in the set of synthetic peptides) were counted as false positives. Subsequently, the true FDR was calculated as the number of these false positives divided by the number of all identifications. For the calculation of the true FLR, we counted peptides with occupied sites that are different from the synthesized ones as false positives, and subsequently the true FLR was calculated as the number of these false positives divided by the number of all identified phosphopeptides.

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

Data availability

Phosphoproteomic data were downloaded from public repositories PXD012174 (ref.⁴), JPST000685 (ref.²⁷), JPST000703 (ref.²⁸), PXD013868 (ref.²⁹), PXD014865 (ref.³⁰), PXD015050 (ref.³¹) and PXD000138 (ref.⁶²). In the case of cross-linking data, files were downloaded from public repositories JPST000916 (ref.³³), JPST000845 (ref.³³) and PXD012723 (ref.³⁴).

Code availability

An open-source implementation with command-line instructions is publicly available (under MIT licence) at <https://gitlab.com/dacs-hpi/AHLF> (ref.⁶³) and includes four independently trained models. In addition, the code repository allows fine-tuning of AHLF as well as training from scratch on third-party or user-specific data.

Received: 27 March 2021; Accepted: 17 February 2022;

Published online: 4 April 2022

References

- Vizcaíno, J. A. et al. A community proposal to integrate proteomics activities in ELIXIR. *F1000Res.* <https://doi.org/10.12688/f1000research.11751.1> (2017).
- Abersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
- Nesvizhskii, A. I., Vitek, O. & Abersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).
- Ochoa, D. et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
- Linding, R. et al. Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415–1426 (2007).
- Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
- Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast open modification spectral library searching through approximate nearest neighbor indexing. *J. Proteome Res.* **17**, 3463–3474 (2018).
- Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. Preprint at *bioRxiv* <https://doi.org/10.1101/483263> (2022).
- Devabhaktuni, A. et al. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat. Biotechnol.* **37**, 469–479 (2019).
- Lu, B., Ruse, C. I. & Yates, J. R. Colander: a probability-based support vector machine algorithm for automatic screening for CID spectra of phosphopeptides prior to database search. *J. Proteome Res.* **7**, 3628–3634 (2008).
- Dorl, S., Winkler, S., Mechtler, K. & Dorfer, V. PhoStar: identifying tandem mass spectra of phosphorylated peptides before database search. *J. Proteome Res.* **17**, 290–295 (2018).
- Zolg, D. P. et al. ProteomeTools: systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Mol. Cell. Proteomics* **17**, 1850–1863 (2018).
- Potel, C. M., Lemeer, S. & Heck, A. J. R. Phosphopeptide fragmentation and site localization by mass spectrometry: an update. *Anal. Chem.* **91**, 126–141 (2019).
- Giese, S. H., Fischer, L. & Rappsilber, J. A study into the collision-induced dissociation (CID) behavior of cross-linked peptides. *Mol. Cell. Proteomics* **15**, 1094–1104 (2016).
- Skinner, O. S. & Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **33**, 717–718 (2015).
- Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
- Kentsis, A. et al. Urine proteomics for profiling of human disease using high accuracy mass spectrometry. *Proteomics Clin. Appl.* **3**, 1052–1061 (2009).
- Cvetesic, N. et al. Proteome-wide measurement of non-canonical bacterial mistranslation by quantitative mass spectrometry of protein modifications. *Sci. Rep.* **6**, 28631 (2016).
- Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
- Avsec, Ž. et al. The Kipoi repository accelerates the community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
- Zhou, X.-X. et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697 (2017).
- Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
- Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl Acad. Sci. USA* **114**, 8247–8252 (2017).
- Tran, N. H. et al. Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nat. Mach. Intell.* **2**, 764–771 (2020).
- Qiao, R. et al. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nat. Mach. Intell.* **3**, 420–425 (2021).
- Xu, L. L., Young, A., Zhou, A. & Röst, H. L. Machine learning in mass spectrometric analysis of DIA data. *Proteomics* **20**, e1900352 (2020).
- Koide, E. et al. Regulation of photosynthetic carbohydrate metabolism by a Raf-like kinase in the liverwort *Marchantia polymorpha*. *Plant Cell Physiol.* **61**, 631–643 (2020).
- Li X. et al. Protein phosphorylation dynamics under carbon/nitrogen-nutrient stress and identification of a cell death-related receptor-like kinase in *Arabidopsis*. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2020.00377> (2020).
- Mergner, J. et al. Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature* **579**, 409–414 (2020).
- Fan, Y. et al. Phosphoproteomic analysis of neonatal regenerative myocardium revealed important roles of checkpoint kinase 1 via activating mammalian target of rapamycin C1/ribosomal protein S6 kinase b-1 pathway. *Circulation* **141**, 1554–1569 (2020).
- Raghuram, V. et al. Protein kinase A catalytic- α and catalytic- β proteins have non-redundant regulatory functions. *Am. J. Physiol. Renal Physiol.* **319**, F848–F862 (2020).
- Giese, S. H., Sinn, L. R., Wegner, F. & Rappsilber, J. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nat. Commun.* **12**, 3237 (2021).
- Lenz, S. et al. Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Nat. Commun.* **12**, 3564 (2021).
- Horn, V. et al. Structural basis of specific H2A K13/K15 ubiquitination by RNF168. *Nat. Commun.* **10**, 1751 (2019).
- Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st Conference on Neural Information Processing Systems* (eds von Luxburg, U. et al.) 4768–4777 (Curran Associates, 2017).
- Janizek, J. D., Sturmfels, P. & Lee, S.-I. Explaining explanations: axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.* **22**, 1–54 (2021).

37. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
38. Yu, F. et al. Identification of modified peptides using localization-aware open search. *Nat. Commun.* **11**, 4065 (2020).
39. Fermin, D., Walmsley, S. J., Gingras, A.-C., Choi, H. & Nesvizhskii, A. I. LuciPHOR: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol. Cell. Proteomics* **12**, 3409–3419 (2013).
40. Fermin, D., Avtonomov, D., Choi, H. & Nesvizhskii, A. I. LuciPHOR2: site localization of generic post-translational modifications from tandem mass spectrometry data. *Bioinformatics* **31**, 1141–1143 (2015).
41. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
42. Bai, S., Koltner, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. Preprint at <http://arxiv.org/abs/1803.01271> (2018).
43. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. Preprint at <https://arxiv.org/abs/1511.07122> (2015).
44. DeGnove, J. P. & Qin, J. Fragmentation of phosphopeptides in an ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **9**, 1175–1188 (1998).
45. Xu, C. & Ma, B. Complexity and scoring function of MS/MS peptide de novo sequencing. In *Proc. Computational Systems Bioinformatics Conference Csb2006* Vol. 4 (eds Markstein, P. & Xu, Y.) 361–369 (World Scientific Publishing, 2006).
46. Kreitzberg, P. A., Bern, M., Shu, Q., Yang, F. & Serang, O. Alphabet projection of spectra. *J. Proteome Res.* **18**, 3268–3281 (2019).
47. Pourshahian, S. & Limbach, P. A. Application of fractional mass for the identification of peptide-oligonucleotide cross-links by mass spectrometry. *J. Mass Spectrom.* **43**, 1081–1088 (2008).
48. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
49. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
50. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at <https://arxiv.org/abs/1412.6980> (2014).
51. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
52. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
53. Bittremieux, W. spectrum utils: a Python package for mass spectrometry data processing and visualization. *Anal. Chem.* **92**, 659–661 (2020).
54. Mendes, M. L. et al. An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).
55. Renard, B. Y. et al. When less can yield more—computational preprocessing of MS/MS spectra for peptide identification. *Proteomics* **9**, 4978–4984 (2009).
56. Goloborodko, A. A., Levitsky, L. I., Ivanov, M. V. & Gorshkov, M. V. Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrom.* **24**, 301–304 (2013).
57. Levitsky, L. I., Klein, J. A., Ivanov, M. V. & Gorshkov, M. V. Pyteomics 4.0: five years of development of a Python proteomics framework. *J. Proteome Res.* **18**, 709–714 (2019).
58. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. Preprint at <https://arxiv.org/abs/1605.08695> (2016).
59. Hulstaert, N. et al. ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. *J. Proteome Res.* **19**, 537–542 (2020).
60. Vaswani, A. et al. Attention is all you need. In *Proc. 31st Conference on Neural Information Processing Systems* (eds von Luxburg, U. et al.) 6000–6010 (Curran Associates, 2017).
61. van den Oord, A. et al. WaveNet: a generative model for raw audio. Preprint at <https://arxiv.org/abs/1609.03499> (2016).
62. Marx, H. et al. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **31**, 557–564 (2013).
63. Altenburg, T. dacs-hpi/AHLF (v1.0.0). *Zenodo* <https://zenodo.org/record/5520955> (2021).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (grant no. RE3474/2-2 to B.Y.R.), by the International Max Planck Research School for Biology AND Computation (to T.A.) and by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (grant nos. 031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A and 031A532B). We thank P. Benner, T. Loka and E. Y. Yuu for proofreading this manuscript.

Author contributions

B.Y.R. conceived the research idea. T.A. developed the deep learning models, carried out model interpretation and performed data analysis. S.H.G. and S.W. contributed to data analysis. T.A., S.H.G., T.M. and B.Y.R. wrote the paper. B.Y.R. and T.M. supervised the research project.

Funding

Open access funding provided by Universität Potsdam.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s42256-022-00467-7>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-022-00467-7>.

Correspondence and requests for materials should be addressed to Bernhard Y. Renard.

Peer review information *Nature Machine Intelligence* thanks Rui Qiao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

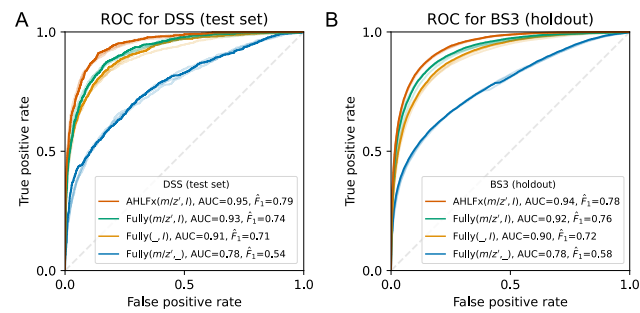
Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022



Extended Data Fig. 1 | AHLF detects cross-linked peptides via transfer learning. A: evaluation on the test set for the DSS cross-linker and **B:** holdout set for the BS3 cross-linker. Receiver operating characteristic of AHLFx (orange) compared to three fully connected networks as baselines (green, yellow and blue) and a random baseline with ROC-AUC=0.5 (dashed line). For each, the top-4 training outcomes are shown and the best is highlighted respectively. For each best model, ROC-AUC and optimal F1-score are summarized in the legends.

	ROC-AUC	\hat{F}_1	Precision (at Recall of 0.95)
AHLFx	0.88	0.64	0.34
Fully(m/z',l)	0.87	0.71	0.20
Fully(-,l)	0.87	0.68	0.22
Fully(m/z',-)	0.72	0.58	0.19

bold indicates the best model performance.

Extended Data Table 1 | AHLFx detects spectra of cross-linked peptides as validated on PXD012723. ROC-AUC, F1-score and Precision at Recall of 0.95 are shown for AHLFx and fully connected networks as baselines.

Data set	Fragmentation/MA/CE	min. Andromeda Score	Bacc (qeAHLFp PhoStar)	F1-score (qeA. P.)	ROC-AUC (qeA. P.)
PXD013868	HCD/FTMS/25	>40	0.933 0.935	0.967 0.969	0.985 0.983
PXD013868	HCD/FTMS/25	>140	0.974 0.963	0.993 0.985	0.998 0.996
PXD013868	HCD/FTMS/25	>200	0.980 0.975	0.997 0.995	0.999 0.998

bold indicates the best model performance.

Extended Data Table 2 | Performance of an instrument-specific AHLFp model that was fine-tuned for Q Exactive data. Metrics are based on predictions for PSMs that have the stated minimum Andromeda score, that is higher Andromeda scores reflect smaller labeling errors.

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 our [Editorial Policies](#) 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 not applicable since no new data was acquired.

Data analysis AHLF is open source under MIT license and available at: [gitlab.com/dacs-hpi/AHLF](https://github.com/dacs-hpi/AHLF)
Python 3.8, tensorflow >2, pyteomics 4.3, SHAP 0.35, path-explain 0.05, scikit-learn 0.23, scipy 1.5, ms-deisotope 0.0.21, ThermoRawFileParser 1.2, MSFragger 3.3, Percolator 3.5, LuciPhor2 2.1 and Ascore (in pyOpenMS 2.7).

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 and 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

our research project is based on the following previously published repositories with accessions numbers (references are in the main text): PXD012174, JPST000685, JPST000703, PXD013868, PXD014865, PXD015050, JPST000916, JPST000845, PXD012723 and PXD000138.

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](https://www.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	<input type="text" value="not applicable since no new data was acquired."/>
Data exclusions	<input type="text" value="not applicable since no new data was acquired."/>
Replication	<input type="text" value="not applicable since no new data was acquired."/>
Randomization	<input type="text" value="not applicable since no new data was acquired."/>
Blinding	<input type="text" value="not applicable since no new data was acquired."/>

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	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging