Made available by Hasselt University Library in https://documentserver.uhasselt.be

CPred: Charge State Prediction for Modified and Unmodified Peptides in Electrospray Ionization

Peer-reviewed author version

VILENNE, Frédérique; AGTEN, Annelies; APPELTANS, Simon; Ertaylan, Gokhan & VALKENBORG, Dirk (2024) CPred: Charge State Prediction for Modified and Unmodified Peptides in Electrospray Ionization. In: Analytical chemistry (Washington), 96 (36) , p. 14382 -14392.

DOI: 10.1021/acs.analchem.4c01107
Handle: http://hdl.handle.net/1942/44276

Title

CPred: Charge State Prediction for Modified and Unmodified Peptides in Electrospray Ionization

Full names

Frédérique Vilenne* [1, 2]
Dr. Annelies Agten [1]
Dr. Simon Appeltans [1]
Dr. Gökhan Ertaylan [2]
Prof. Dr. Dirk Valkenborg [1]

[1]: Hasselt University, Data Science Institute, Hasselt, Limburg, BE 3500
[2]: Flemish Institute for Technological Research, Health department, Mol, Antwerpen, BE 2400

Formatted: Dutch (Netherlands)

Abbreviations

AA = Amino Acids
CID = Collision-Induced Dissociation
DL = Deep Learning
ETD = Electron Transfer Dissociation
LC = Liquid Chromatography
LSTM = Long Short-Term Memory
MS = Mass Spectrometry
m/z = Mass to Charge
PRIDE = Proteomics Identifications Database
PCC = Pearson correlation coefficient
PTMs = Post-Translational Modifications
ReLU = Rectified Linear Unit
RNN = Recurrent Neural Networks

Keywords

Mass spectrometry
Proteomics
Charge state
Deep learning
Neural network

Total number of words

8092 words in total

Abstract

The mass-to-charge ratio serves as a critical parameter in peptide identification via mass spectrometry, enabling the precise determination of peptide masses and facilitating their differentiation based on unique charge characteristics, especially when peptides are ionised by tools like electrospray ionisation, which produces multiply charged ions. We developed a neural network called CPred, which can accurately predict the charge state distribution from +1 to +7 for modified and unmodified peptides. CPred was trained on the large-scale synthetic training data, consisting of tryptic and non-tryptic peptides, and various fragmentation methods. The model was further evaluated on independent, external test datasets. Results were evaluated through the Pearson correlation coefficient and showed high correlations up to 0.9997117 between the predicted and acquired charge state distributions. The effect of specifying modifications in the neural network and feature importance was further investigated, revealing the value of modifications and vital peptide properties in holding on to protons. CPreds' accurate predictions of the charge state distribution can play an essential role in boosting confidence in peptide identifications during rescoring as a novel feature.

Statement of Significance

Mass spectrometry-based proteomics is continuously advancing, both from the experimental and the computational point-of-view, yielding higher-quality spectra. From the bioinformatics angle, analysing these spectra is becoming easier and harder at the same time. State-of-the-art algorithms are continuously being developed to achieve the highest number of peptide spectrum matches. In doing so, the databases often used to acquire these identifications keep growing, possibly leading to a higher number of false identifications. To control the false discovery rate, rescoring is becoming almost mandatory. Algorithms such as Percolator use various features, such as the retention time, to estimate the false discovery rate. One key element used to identify peptides is the mass-to-charge ratio, and a potential feature to facilitate rescoring is the probability of a peptide occurring in said charge state. CPred was developed especially for this reason, predicting the charge state distribution for (un)modified peptides with high accuracy. The algorithm is released as an easily applicable Python package, ready to be incorporated in bioinformatics pipelines, boosting confidence in identifications.

Introduction

The charge state of a peptide originates from ionising the peptide, usually after the initial separation of the mixture by liquid chromatography (LC). A wide selection of ionisation techniques are available. However, the most common methods are electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI) [1, 2]. Both techniques are so-called soft ionisation techniques, allowing ionisation with little fragmentation of the molecules. In ESI, the analyte solution is pumped through a small needle under high voltage. At the tip of the needle, the voltage causes the formation of an aerosol of charged droplets. The charge on the aerosol is a direct result of the strong electric field at the tip of the needle. Next, the aerosol moves towards the inlet of the mass spectrometer, where it evaporates due to the heat and reduced pressure in the vacuum chamber of the mass spectrometer. This causes the droplets in the aerosol to reduce in size and increase in charge density. As the aerosol evaporates, the remaining charges on the droplets repel each other, leading to a Coulombic fission of the droplets. This leads to the analyte molecules within the droplets being released in a gas phase as ions. These ions can carry one or multiple charges. These ions are further introduced in the mass spectrometer for mass analysis. ESI has several advantages, such as the ability to analyse large molecules, producing multiple charged ions and minimal fragmentation of molecules [3]. On the contrary, MALDI primarily produces singly charged ions.

The charge state has multiple important roles. The most obvious example here is the isotope distribution of the precursor ion. Different charge states lead to a different representation of the isotope distributions of the same precursor ion. Secondly, the charge state also influences the fragmentation pattern of the precursor ion in MS^2 spectra. A study from 1998 by Downward et al. researched the effect of the charge state on the fragmentation behaviour through collision-induced dissociation (CID). Their findings indicated that the fragmentation behaviour of doubly charged ions was influenced by the localisation of basic amino acids (AA). Additionally, they found that the fragmentation pattern may differ between singly and multiply charged ions. For a more extensive overview of all their findings, we refer to the original article [4]. This fragmentation behaviour of peptides in CID can be explained by the mobile proton model of Vicki Wysocki and Simon Gaskell. To briefly summarise the model, protons can move around the molecule and influence how the peptide fragmentation[5, 6]. Other research mentions different a fragmentation method, electron transfer dissociation (ETD), favouring ions with higher charge states [7].

As the importance of the charge state is widely acknowledged, researchers have attempted to accurately determine the charge state of peptides and proteins for decades. In 1988, Covey et al. found that the highest charge of an ion was correlated with the number of basic residues in the ion[8]. Several years later, in 1991, Smith et al. compiled a list with the highest charge state for peptides and proteins, combined with the number of basic residues in the ion, confirming this relationship, albeit with notable deviations [9]. In the following years, several solution-phase equilibrium models and gas-phase basicity models were proposed to explain the charge state in ESI [10-12]. More recently, thanks to computational advancements, various modelling frameworks have been developed to predict the charge state and distribution. In 2008, Charger was introduced to predict the charge state given an MS^2 spectrum [13]. In the following years, Charge Prediction Machine (CPM), ETDz, and a modelling framework by Liu et al. were released [14-16]. All the tools mentioned above were specifically designed to predict the charge state for spectra acquired through ETD fragmentation and varying ionisation methods. While each tool performed well in its respective settings, they all suffered from some drawbacks. They were all limited to a small amount of training data. Additionally, the modelling framework by Lui et al. was limited to predicting up to a charge state of +3. CPM and ETDz were able to predict charge states up +7. More recently, Guan et al. developed a tool to predict the probability of a charge state, given a peptide sequence through a deep learning (DL) model [17]. Guan et al. leveraged a deep learning framework using Long Short-Term Memory (LSTM) layers, predicting the charge state up to +5. Given a peptide sequence, they limited themselves to a small set of post-

translational modifications (PTMs) and tryptic peptides. This is a drawback of using the current model, given the interest in PTMs and their crucial role in the bioactivity of peptides. Another downside of their methodology was validating their model on a holdout dataset, while an external dataset would give a more accurate representation of the model's predictive capabilities.

During this study, we created a DL framework to predict the charge state probability of a given peptide sequence with PTMs. Through feature engineering, the model is capable of accurately predicting the charge state distribution for any peptide sequence, both tryptic and non-tryptic. Additionally, the methodology allows for the presence of any modification in the UNIMOD database, even when unseen during training [18]. The complete ProteomeTools project was leverages as a readily available, sufficiently large, high-quality dataset containing data of tryptic and non-tryptic, modified and unmodified peptides, ionised using ESI [19]. By utilising this data, the model will be widely applicable as a supportive component for the current state-of-the-art proteomics identification algorithms for qualitative and quantitative proteomics, including immunopeptidomics. In addition, insights into the charging mechanism of peptides is acquired through investigation of the feature importance.

Experimental Section

Data

The model was trained using the complete ProteomeTools dataset. For a detailed description of the experimental settings of the entire ProteomeTools project, we refer to the original ProteomeTools paper [19]. In brief, all peptides were synthetically manufactured and divided into pools of approximately 1,000 peptides per pool. Each peptide pool consisted of peptides with different masses wherever possible to simplify identifications. All peptide pools were subjected to an LC-MS/MS experiment with 60 minutes of High-Performance LC, followed by MS/MS on an Orbitrap Fusion Lumos (Thermo Fisher Scientific). All data were analysed using MaxQuant (version 1.5.3.30) with peptide pool-specific databases with a false discovery rate of 0.01 [20]. A more detailed description of the database search of each specific dataset can be found in their respective articles or on the PRIDE repository. All raw data and search results are publicly available on the Proteomics Identifications database (PRIDE) [21], accessible through the identifiers PXD004732, PXD010595, PXD021013, PXD023119, PXD023120, and PXD009449. ProteomeTools consists of qualitative data (PXD004732, PXD010595, and PXD021013) and quantitative data (PXD023119 and PXD023120) with a TMT 6-plex label (Thermo Fisher Scientific). The datasets PXD021013 and PXD023120 consist of immunopeptidomics data and products of the digestion enzymes AspN and LysN, with PXD23120 being TMT-labelled. Lastly, PXD009449 is a smaller-scale dataset from ProteomeTools consisting of 5,000 peptides with 21 commonly occurring PTMs. The search results from each dataset were downloaded, and all identifications were used for further processing. Each uniquely identified peptide's proportions for each charge state from +1 to +7 were computed, accounting for PTMs. This processing resulted in 5,861,945 unique (un)modified peptides. The data was filtered to reduce the sheer dataset size for training and get more accurate proportions per charge state so that each unique (un)modified peptide had to be detected at least ten times. Note that this choice was arbitrary and resulted in 2,769,900 remaining training observations.

For evaluating the predictive model, an independent test data set was assembled. To do so, several datasets were used. The first dataset was the multiprotease dataset of Bekker-Jensen et al. [22]. The dataset contains HeLa cells digested by trypsin, Lys-C, Glu-C and chymotrypsin measured on an Orbitrap, fragmented by CID and HCD. A second dataset was created by Davis et al. [23], containing proteins originating from MCF-7 breast cancer cells. Samples were digested by trypsin and elastase, measured on an Orbitrap Fusion, and contained peptides fragmented by CID and HCD. Both datasets were also analysed using MaxQuant (version 1.5.3.30). Charge state +1 was excluded from the experiment, which could influence the neural network's performance on the test data set. We refer to the respective papers for further details on the experiment itself. A more detailed description of

the data analysis of the datasets can be found in the original paper of Prosit [24]. Similarly to the ProteomeTools data, observations were filtered to be detected at least ten times.

Feature Engineering

Given a peptide sequence and PTMs, features are generated through a manually programmed Python workflow (version 3.9.13) [25]. A complete list of the sixty-nine features that were developed based on the given information, is provided in Table S1. We made a distinction between static and sequential features. The static features are summary statistics derived from the peptide sequence. The sequential features are the one-hot encoded peptide sequence, the isoelectric point of each AA in the peptide sequence and the hydrophobicity of each AA in the peptide sequence. The hydrophobicity index determined by Black et al. was used [26]. We used the isoelectric point and hydrophobicity index of the raw peptide sequence, not accounting for the PTMs because the influence of all PTMs on these indices remains unknown at the time. The isoelectric point and hydrophobicity index were included as they may influence proton binding and thus acquire higher charge states. Furthermore, to account for variable peptide sequence lengths, peptide sequences were padded to a length of fifty using zeros.

To account for the (unseen) PTMs, we followed a similar approach as DeepLC [27]. A widespread practice of encoding peptide sequences is by one-hot encoding the peptide sequence. However, given the current interest in PTMs and the vast amount of potential PTMs, it is impossible to one-hot encode all possible AAs and their possible modified state. DeepLC approached this issue by one-hot encoding the raw peptide sequence, not accounting for any modifications, and using the elemental composition of the peptide sequence, including the PTMs, as additional input for their model. This was proven to give their retention time model the capability to predict the retention time of unseen modifications. We used the same foundation in our model. We extended it by computing the monoisotopic and average mass of the modified peptide sequence, allowing for all modifications in the UNIMOD database, even when unobserved during training. By including the monoisotopic and average masses of the peptide sequences, accounting for the PTMs, we will enable the model to generalise for PTMs unseen during training.

Model architecture

A multi-input deep neural network architecture was constructed using a TensorFlow and Keras framework [28]. Four distinct input pathways were established for each peptide sequence feature type under consideration. Sequential attributes, such as one-hot encoded peptide sequences, AA isoelectric points, and hydrophobicity indices, were analysed through bidirectional LSTMs[29]. The static features were subjected to regular dense layers. Results from the initial layers were concatenated and further processed by a dense layer and a final dense layer to predict the probability of each peptide occurring in charge state +1 to +7. Dropout and batch normalisation were used to improve the model performance[30]. A more in-depth explanation of the model is available in the supporting information, combined with a visualisation of the model architecture in Figure S3. The models' hyperparameters were optimised using the Hyperband approach on a smaller subset of the ProteomeTools data [31]. The model was trained using a training-validation-test split of the ProteomeTools data with a mean squared error loss function and accuracy as the evaluation metric. The model was further evaluated through the Pearson correlation coefficient (PCC) between the predicted and experimental charge state distribution and the Shannon Entropy as an additional information measure[32]. Lastly, the feature importance of the model was investigated on a 50% sampled subset of the data. Each feature was analysed independently by permutation in 5-fold and comparison to the baseline results through the PCC.

All computations were performed on the Flemish Supercomputer Centre using a Xeon Gold 6140 CPU with 18 cores and 192GB RAM and an NVIDIA P100 SXM2 GPU with 16GB GDDR.

Results and Discussion

The results section is divided into two subsections. Firstly, we performed exploratory research into the training data of ProteomeTools, effects of modifications on the charge state distribution and the consistency of the charge state distribution across different datasets. The second subsection discusses the evaluation of the neural network measured through the PCC and Shannon entropy on the test data. We investigated the effects of modifications on the model's accuracy and evaluated the model's performance on independent test datasets. In addition, the Supporting Information contains a section on the model development, discussing the hyperparameter tuning and final settings of the model.

Data exploration

Firstly, we explored the training data, depicted in Figure 1~~Figure 1~~. Figure 1~~Figure 1~~A illustrates the charge state distribution for the complete ProteomeTools dataset. It indicates that most peptides have a high probability of occurring with a charge state of +2 and a lower likelihood of a charge state of +3. However, it is also visible that based on the outliers, there are peptides present with a high probability of occurrence for other charge states, such as +1 and +4 to +7, indicating the importance of a prediction tool being capable of predicting such a wide range of charge states. The effects of PTMs on the charge state distribution are shown in Figure 1~~Figure 1~~B. Compared to their unmodified counterpart, peptides modified by methyl groups, albeit one, two or three methyl's, cause a shift in the charge state distribution towards a higher charge state. This shift in charge state distribution could be dedicated to the methylation residues increasing the basicity of the side chain, causing it to retain the charges. Somewhat similar findings were reported by Krusemark et al. [33]. They reported small shifts in the charge state distribution towards higher charge states due to methylation in proteins. However, it should also be noted that the complete opposite was reported by Zolg et al. in the original ProteomeTools paper regarding the PTM dataset [34]. They found that the methylation did not cause any shifts in charge state distribution. However, it should be noted that they only looked towards the results of HCD fragmentation with 28% collision energy, which could contribute to the discrepancy between our findings. A similar effect is observed for GlyGlycylation, a PTM where two glycine residues remain at a ubiquitinated Lysine residue. Our findings indicate a slight increase in the charge state distribution. At the same time, Zolg et al. reported no changes in the charge state distribution, which could be attributed to the same rationale for the methylation PTMs. An increase in the charge state distribution could be explained by the addition of the nitrogen and oxygen atoms. In particular, nitrogen makes a suitable binding spot for protons because it is more electronegative than Hydrogen, attracting electrons more strongly. As a result, the electrons on the Nitrogen are available for donation, and it can readily accept a proton for a covalent bond. Oxygens can also act as a basic site and accept protons but generally do so to a lesser extent than Nitrogen. A minor shift towards higher charge states in charge state distributions is observed in TMT6-plex labelling, hydroxylation of Proline, Methionine oxidation and carbamidomethylation of Cysteine, depicted as other in Figure 1~~Figure 1~~B. The hydroxylation of Proline and Methionine oxidation may also be explained by the addition of Oxygens. At the same time, the carbamidomethylation of Cysteine contains a methyl group similar to the other methylation PTMs. Lastly, the TMT6-plex increases the charge state of peptides by adding a positively charged tag to the N-terminus of the peptide. The TMT-tag is a bulky molecule that can disrupt the secondary structure of peptides, making them more susceptible to protonation. The TMT-tag also contains a guanidinium group, a strong base that can easily attract protons. The effect of isobaric labelling, including TMT-labelling, on the charge state was previously described in literature by Thingholm et al. [35]. They specifically discussed the increase of charge state in phosphopeptides due to the presence of isobaric labels, leading to a reduced number of identifications [35]. Other modifications remained somewhat constant, with primarily a charge state of +2.

The effect of the fragmentation method was also investigated and elaborated upon. We investigated the charge state distribution for each fragmentation method. The results are visualised in the Figure S1 and S2, together with a short discussion on the results.

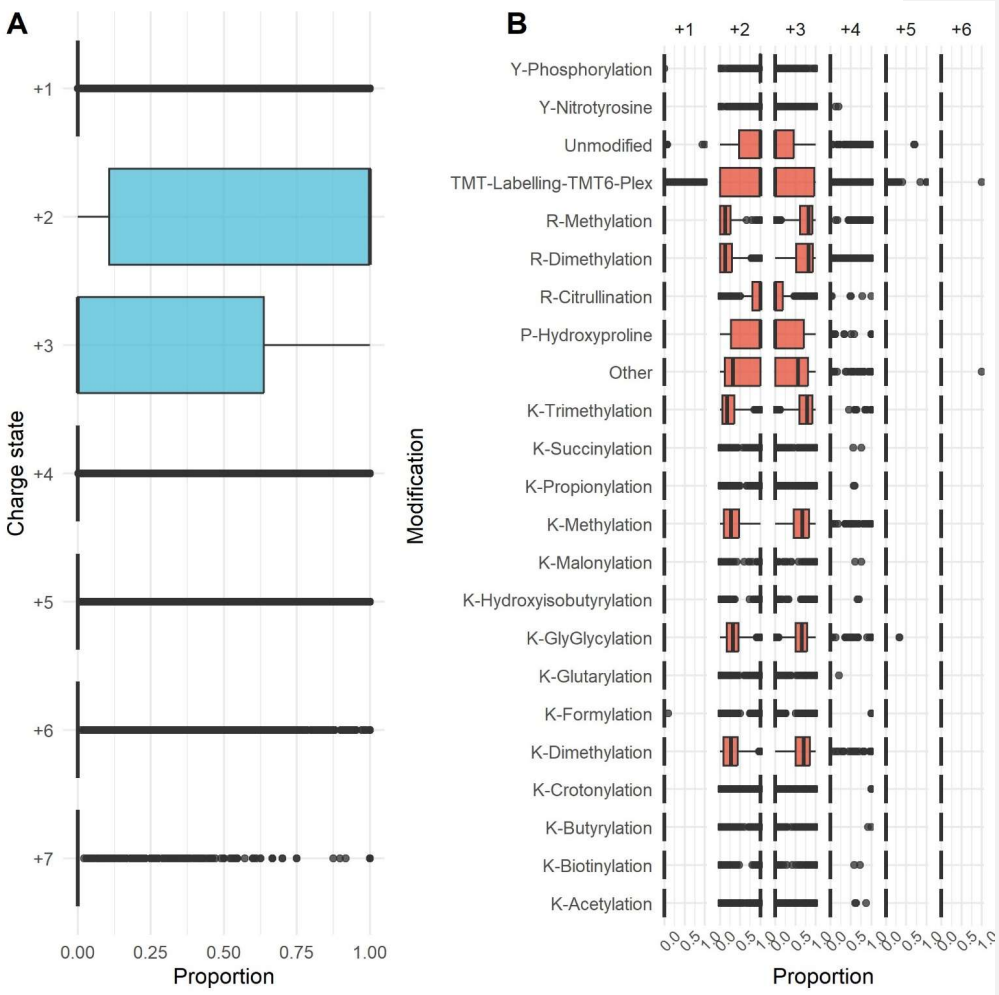

Figure 1 Exploration of the ProteomeTools data. ***(A)*** *The charge state distribution from charge +1 to +7 for the complete ProteomeTools dataset.* ***(B)*** *The charge state distribution for every modification within the ProteomeTools PTM dataset, including TMT6-plex labelling of the quantitative ProteomeTools datasets.*

Finally, we looked into the consistency of the charge state distribution. To do so, all common peptides were selected in the complete ProteomeTools project, Davis dataset and Bekker-Jensen multiprotease tryptic dataset. This yielded a total of 408 common peptides. The pairwise Pearson correlation between each dataset was calculated for charge state +1 to +5. We limited ourselves to charge state +5 because the common peptides did not occur in higher charge states. The results are shown in Figure 2. When looking at the correlation between the ProteomeTools and Davis datasets, a high

correlation close to 1 is observed in the diagonal, indicating a high correlation between the charge state distribution in both datasets. This indicates that the charge state distribution remained relatively constant between all experimental conditions. Even for charge state +1, which rarely occurred, a relatively high Pearson correlation of 0.66 is acquired. When comparing the ProteomeTools dataset and the Davis dataset to the Bekker-Jensen tryptic dataset, the correlations are slightly lower, varying between 0.68 and 0.83 for charge state +2 to +4, the most common charge states. The correlations for charge states +1 and +5 are below 0.5, indicating some dissimilar findings. This could be because these charge states aren't as present in the dataset, which leads to more variability and less stability when computing the Pearson correlation. Another noteworthy observation for each pairwise comparison is the high negative correlation between charge states +2, +3 and +4. This is attributed, for example, to the fact that when the probability of charge state +2 increases, the probabilities for charge state +3 or +4 strongly decrease. A noteworthy conclusion for these findings is the remarkable stability of the charge state distribution across different experiments. Despite the general similarity of most proteomics experiments, slight variations were observed based on specific experimental conditions. These included the use of different buffers (Dimethyl Sulphoxide for ProteomeTools and Davis, Tris-HCl for Bekker-Jensen, and phosphate-buffered saline for Davis), detergents (sodium dodecyl sulfate for Davis), alkylating agents (iodoacetamide for Davis and chloroacetamide for Bekker-Jensen), reducing agents (dithiothreitol for Davis and tris(2-carboxyethyl)phosphine for Bekker-Jensen), and mass spectrometers (Orbitrap Fusion Lumos for ProteomeTools and Davis, Q-Exactive HF for Bekker-Jensen). Other experimental parameters, such as the mobile phase and LC type, were highly similar across all experiments. To our knowledge, these findings have not been previously reported in the literature.

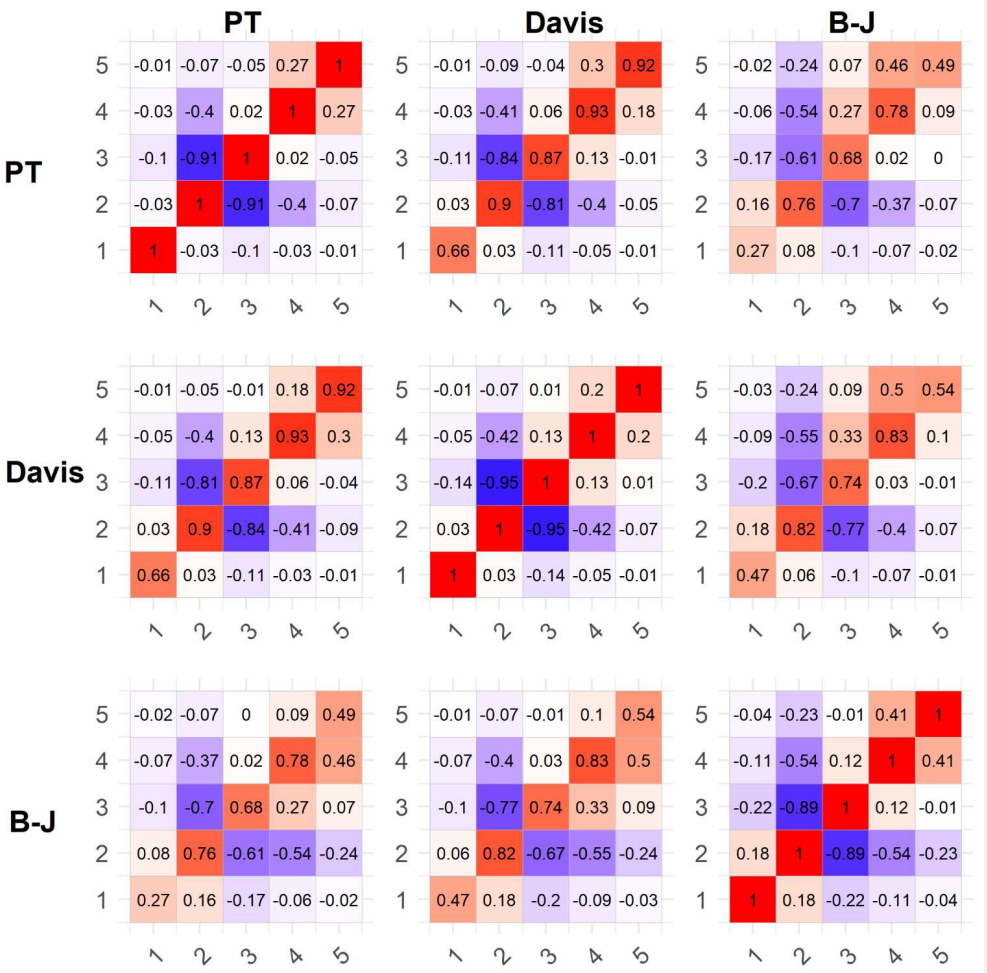

Figure 2 Pairwise correlation between common peptides in the ProteomeTools (PT) project, Davis dataset and Bekker-Jensen tryptic dataset (B-J). Red indicates a positive correlation, blue indicates a negative correlation and white indicates no correlation. The correlation itself is shown in the cells. The x- and y-axis denote charge states +1 to +5.

Model evaluation

The neural network was further evaluated by predicting the charge state distribution on the test-split of the ProteomeTools dataset. We created a confusion matrix to see if the charge state with the highest predicted probability corresponds to the charge state with the highest experimental probability for each peptide. Table 1 shows the results of these predictions. In general, an accuracy of 0.9379 was acquired, indicating the model is capable of predicting the most abundant charge state in 93.79% of the cases. It is noteworthy that the accuracy is not entirely suitable when working with imbalanced classes, which is the case for the test data. Most of the observations had the highest

probability of occurring in charge state +2 and +3, similarly as observed in Figure 1~~Figure 1~~. A trend may be observed that for experimental charge states higher than +4, the model tends to classify the charge state one charge lower. This is most likely a consequence of the entire ProteomeTools data, including the training data, predominantly containing peptides with a higher probability of carrying two or three charges. Unquestionably, when looking towards the higher charge states such as +5 to +7, the predictions tend to be mostly incorrect, most likely due to a sampling bias as most peptides where mostly abundant at charges +2 and +3. We further investigated peptides who were mostly abundant for charge states +5 to +7. The main functionality of CPred is the prediction of the charge state distribution, as such, we looked at the PCC of the experimental and predicted charge state distribution of these peptides. CPred acquired a PCC of 0.9150 for the peptides who are mostly abundant from charge states +5 to +7. Depending on the importance of these charge states, it could prove to be beneficial to retrain the model in the future on additional data containing highly charged ions. Noteworthy, for the most abundant charge states, +2 to +4, the model predicted the most abundant charge state of the peptides correctly. For completeness, we provided Table S2, which provides summary statistics per class such as the sensitivity, specificity, prevalence, positive and negative predictive value. As the prediction of the most abundant charge state is not the scope of CPred, we will not discuss it and only provide it as additional information.

Table 1 Contingency table of the CPred predictions on the ProteomeTools test data. The most abundant experimental charge state is compared with the most abundant predicted charge state.

Charge state		**Charge state with highest abundance (experimentally)**						
		+1	+2	+3	+4	+5	+6	+7
Charge state with highest abundance (predicted)	+1	20.840	1.148	4	0	0	0	0
	+2	1.620	402.009	9.783	82	1	0	0
	+3	10	19.820	182.196	2.744	6	2	0
	+4	0	250	5.293	28.112	555	18	0
	+5	0	5	77	497	1.197	53	1
	+6	0	0	1	10	22	75	13
	+7	0	0	0	2	0	0	2

While the classification of charge states is an interesting topic to investigate, the main goal of the research was creating a model capable of accurately predicting the charge state distribution. We used the PCC to see if the predicted charge state distribution was similar to the experimental charge state distribution. Figure 3~~Figure 3~~A shows the distribution of the PCC. The Pearson correlation takes on values between -1 and 1, where 1 indicates a strong positive correlation or highly similarity between the predicted and experimental distribution. The plot indicates that most predictions acquired a high value of the PCC, which suggests that the predicted charge state distribution is highly similar to the experimentally obtained charge state distribution. Furthermore, a median PCC of 0.9997117 was acquired, reassuring the high-quality predictions of the neural network. While comparison is not truly justifiable, which will be discussed in-depth at the end of the Results and Discussion, the model by Guan et al. acquired a median PCC of 0.997. As an additional evaluation method for the neural network, we calculated the Shannon entropy of the experimental charge state distribution in function of the PCC, as shown in Figure 3~~Figure 3~~B. The entropy is a measure explaining the randomness of the charge state distribution. High values, up to a binary logarithm of 7, for the entropy indicate that the probabilities are evenly distributed across the categories. In CPreds' case, the entropy can take on values between 0 and 2.807355. A low value for the entropy indicates that specific outcomes are more likely than others, making the distribution predictable. When a particular charge state has a probability of 1 and all other charge states have probabilities of 0, an entropy of 0 is acquired. The ProteomeTools data mainly consisted of peptides with a low value for the entropy, indicating that they were primarily detected with a single charge state. This is an indicator that most peptides are relatively stable with

regards to holding on to charges, as they only occurred in a sole charge state. Ideally, in an experimental setting, the charge state of peptides would remain constant. A cloud of observations towards an entropy of 0.7 and a second cloud towards an entropy of 1.1 can be observed. There is no clear relationship between acquiring a lower PCC and high values for the entropy. This indicates that the neural network can easily predict the charge state distribution for peptides with a more complicated charge state distribution. As to the pattern witnessed, further investigation was shown in Figure S4. The flat line at 0 Entropy is caused by peptides occurring at a single charge state with a probability of 1, which amounted to approximately 70% of the holdout data. The increase towards an entropy of 0.7 is predominantly the result of peptides partially composing of a charge state +2, +3 and slightly +4. The increasing entropy towards 1.1 consists of peptides partly belonging to charge states +2 to +4 and partially +5. Peptides occurring in multiple charge states exhibit more complex charge state distributions, nevertheless, the model achieved a high a PCC for these peptides. Lastly, Figure 3~~Figure 3~~C depicts a boxplot to visualise the distribution of probabilities for each peptide to occur in a certain charge state for both the predicted and experimentally acquired charge states. The high similarity in the shape of both distributions, including the outliers, suggests the neural network accurately predicts the charge state distribution. Notably, for charge state +7, discrepancies can be observed, where experimentally observed charge states had outliers with a probability of 1, which the neural network failed to predict this. Similarly to the findings in Table 1, this discrepancy might indicate a slight bias towards lower charge states within the neural network, which can be addressed, if required, by retraining CPred with highly charged peptides.

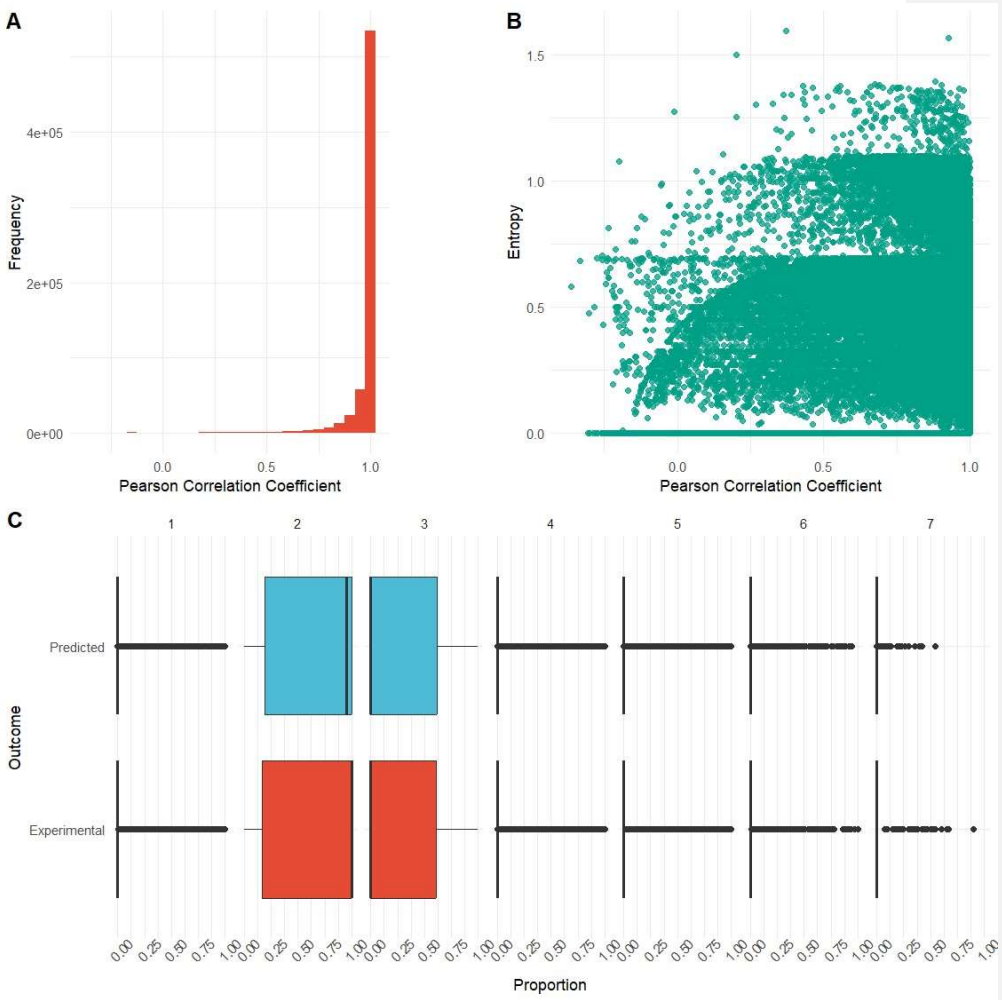

Figure 3 Results of the Neural Network. ***(A)*** *The distribution of the Pearson Correlation Coefficient was calculated between the predicted charge state distribution and the experimentally acquired charge state distribution.* ***(B)*** *The Shannon entropy of the experimentally obtained charge state distribution in function of the Pearson correlation coefficient between the predicted and experimentally acquired charge state distribution.* ***(C)*** *The distribution of probabilities for each charge state is visualised for both the predicted charge states in blue and the experimental charge states in red.*

A permutation-based approach was used to investigate each feature separately to gain insight into the most influential factors in determining the charge state distribution. The baseline PCC was computed based on a random sample of 50.000 observations. Subsequently, each feature underwent permutation, followed by the recalculation of the PCC. The difference between the baseline PCC and the permuted PCC was calculated. This iterative process was repeated five times, and the average influence per feature was calculated. The eight most influential features are visualised in Figure

4~~Figure 4~~. Notably, the sequential feature representing the isoelectric point per AA was shown to be the most influential feature in the neural network. This importance may be explained due to the basic AA side chains easily being protonated. The LSTM layers embedded within the neural network can capture regions within the sequence that are more basic, thereby increasing the likelihood of retaining protons. The importance of the isoelectric point per AA was further underscored during the hyperparameter tuning. The Hyperband parameter optimization process allocated the largest amount of units to the bidirectional LSTM layers, responsible for analysing this particular feature. Additionally, both average mass and monoisotopic mass were shown to be significant contributors to the neural network's predictive performance. The relationship between peptide mass and length follows a linear pattern, where longer peptides can hold on to charges easier in comparison to small peptides due to a higher number of binding sites. Additionally, heavier peptides require higher charge states to be detectable within the m/z range of an MS device. Lastly, the PTMs were incorporated into the mass statistics, impacting the charge state distribution. The importance of the 13-carbon and 15-nitrogen elements can be attributed to the presence of the component in the TMT6-plex, contributing to higher charge states in general as observed in Figure 1~~Figure 1~~ and as discussed in the research of Thingholm et al. [35]. Lastly, the one-hot encoded sequences showed less influence, possibly because the isoelectric point of each AA carried the same information except on a higher level due to the isoelectric point. Other features, including the hydrophobicity index per AA, were shown to be less influential within the neural network.

Guan et al. relied solely on the AA sequence to predict the charge state distribution, whereas CPred incorporates a broader array of features. Notably, for CPred, the one-hot encoded sequences did not emerge as the most important feature. An interesting topic for further investigation would be the performance of their model when incorporating the iso-electric point per AA instead of one-hot encoded sequences.

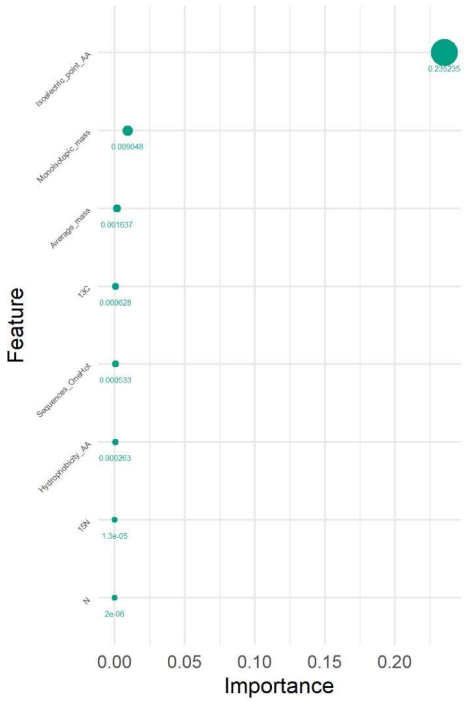

Figure 4 Feature importance. On the y-axis are the top 8 most influential features. The importance on the x-axis depicts the change in the Pearson correlation coefficient when subtracting the mean permuted Pearson correlation coefficient from the baseline Pearson correlation coefficient.

The research extended to exame the impact of allowing modifications in predicting the charge state distribution. CPred predicted the charge state distribution for the holdout dataset, considering both modified and unmodified scenarios. The PCC were computed between the predicted distributions and the experimentally observed charge state distribution.. The results, presented in Figure 5~~Figure 5~~, indicate that the majority of the modifications have a limited influence on the PCC, consistent with our earlier findings illustrated in Figure 1~~Figure 1~~B. A significant deviation in the PCC is observed for a TMT6-plex, aligning with the trends observed in Figure 1~~Figure 1~~B. Failure to specify TMT6-plex modifications results in a shift towards lower PCC values, suggesting a decreased similarity between the predicted and experimental charge state distribution. These findings are aligned with the previous findings of Thingholm et al. [35] and our previous findings in Figure 4~~Figure 4~~, where permuting 13-carbon and 15-nitrogen noticeably affected the neural network's performance. More minor shifts towards lower PCC values are observed for methylation and glutarylation when PTMs are left unspecified. Although Methylation influences the charge state distribution, its impact is comparatively less pronounced than that of TMT6-plex. Glutarylation, which demonstrated minimal influence in Figure 1~~Figure 1~~B, exhibits a reasonable minor shift in this context. An enhancement to the model could involve incorporating the PTM effect on the iso-electric point per AA, considering it was the most influential feature. This adjustment would more accurately reflect the modified residue's basicity, a critical factor in charge retention as supported by our previous findings.

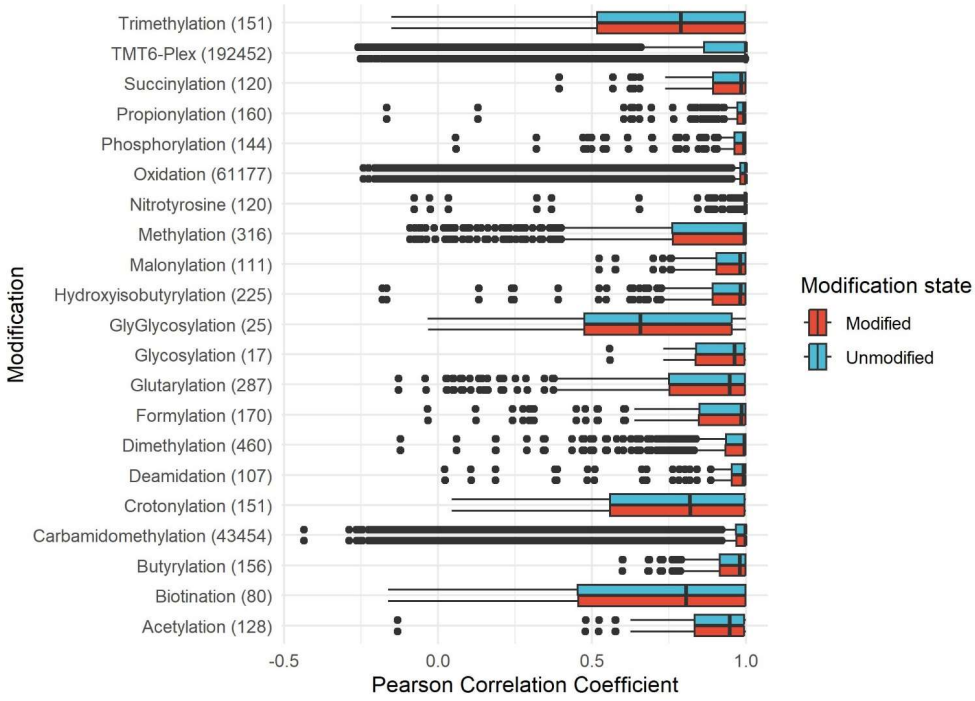

Figure 5 The effect of modifications on the neural networks' performance. The distribution for the Pearson correlation coefficient is shown on the X-axis for both the holdout data where modifications were specified (red) or unspecified (blue). On the Y-axis, all modifications in the ProteomeTools dataset are shown with the number of observations between brackets.

The model was further evaluated on independent test data using the Bekker-Jensen multiprotease and Davis datasets and visualised in Figure 6~~Figure 6~~. The Bekker-Jensen glutamyl endopeptidase, commonly known as Glu-C, and chymotrypsin datasets acquired the lowest median PCC with respectively 0.7884 and 0.7932. The most obvious explanation is that peptides cleaved by chymotrypsin and Glu-C were absent from the training data. The ProteomeTools data consisted of peptides originating from trypsin (cleavage site [RK].<P>), Lys-C (cleavage site [K]), AspN (cleavage site [N-terminal of N]) and immunopeptides. Both chymotrypsin (cleavage site [FWY]) and Glu-C (cleavage site [E]) have different cleavage properties and may be more troublesome to predict for the neural network. The Lys-C and tryptic data from the Bekker-Jensen dataset achieved higher median PCC of 0.8623 and 0.9211. Both proteases were present in the training data and were more predictable for the neural network. Lastly, the Davis dataset, a combination of trypsin and elastase (cleavage site [AG]) acquired a median PCC of 0.9868. Notably, as mentioned before, the experiment of Davis excluded charge state +1, which influenced the PCC in a slight negative matter. In total, the neural network predicted a probability higher than 0.1 only 11 times for charge state +1 in the Davis dataset, making the impact of the experimental settings negligible. Given the results of the Bekker-Jensen Lys-C, Bekker-Jensen tryptic and Davis dataset results, the model is able to accurately predict the charge state distribution, even with regards to different experimental settings. One reason why such varying results are observed for the different datasets and proteases is that the proteases have different cleavage sites, as mentioned between brackets. This causes the proteins to be cut into peptides of different lengths. While the model takes the peptide length into account as a feature, this may still affect the models' accuracy. Additionally, these peptides may also have different charge characteristics compared to regular tryptic peptides. In order to enhance CPreds' predictive capabilities, it would be worthwhile to retrain the model on the protease of choice when the performance may not be sufficient. A final, potential reason for the lower PCC on the Bekker-Jensen datasets when compared to the ProteomeTools test-split and Davis dataset could be associated with variations in experimental conditions. Figure 2 illustrated that the charge state distribution of shared peptides in the Bekker-Jensen dataset differed more significantly from those in the ProteomeTools and Davis datasets. This discrepancy suggests that the peptides exhibited distinct behaviour in terms of charge retention. Consequently, the predictions of charge state distribution generated by CPred may diverge from the experimentally acquired distributions in the Bekker-Jensen datasets, contributing to the observed reduction in similarity between the predicted and experimental outcomes.

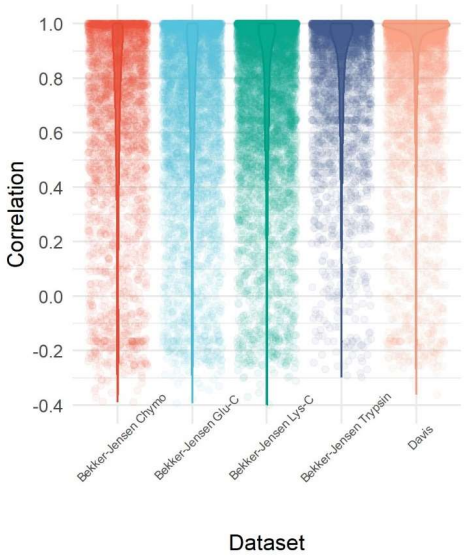

Figure 6 The evaluation of the neural network on independent test data. The X-axis show the different datasets used for evaluation. On they Y-axis the Pearson correlation coefficient. The datasets each have a different colour.

A final point of discussion which was left open until now, was why CPred is not entirely comparable to the model of Guan et al. [17]. The origin of the problem lies in the nature of the training data used for the model. The training data used to develop the model is based on spectral counting, meaning that the charge state distribution was calculated from spectral matches. The number of spectral matches belonging to a certain peptide are influenced by the experimental settings, both on the mass spectrometer and during analyses. A neural network based on ion counting in the MS^1-layer would be more representative to the true underlying charge state distribution of a peptide, in a similar fashion as the algorithm developed by Guan et al. [17]. This was the main reason why the algorithm was not compared to CPred, as both algorithms work in a different fashion. Additionally, the webservice of the model by Guan et al. is no longer available [17].

Concluding remarks

Our research shows that our neural network, called CPred, is capable of accurately predicting the charge state distribution for both unmodified and modified peptides from charge state +1 to +7. The results were discussed in depth during the results section and will only be mentioned here briefly. Peptides predominantly occur within the range of charge state +2 to +4. Most peptides are stable and only occur in a single charge state. The presence of modifications affected the charge state distribution, and specifying the modifications benefitted the model's performance. Data exploration showed that the charge state distribution of peptides remained constant over different experimental conditions, which was not reported in literature before. Throughout our research, the importance of the isoelectric point in retaining charges was shown, unravelling key properties in the charge retention behaviour of peptides. CPred was shown to predict the charge state distribution accurately through the PCC. Additionally, CPred was trained on peptide spectrum matches from various fragmentation methods and proteases. It should be noted however, that the model's performance on tryptic

peptides was significantly better compared to other proteases, making it worthwhile to retrain the model on the protease of interest.

The charge state distribution may be a useful feature during the rescoring of a peptide spectrum match in a similar fashion as the retention time. Having an accurate prediction of this distribution may contribute towards more accurate identifications. As CPred was developed on spectral counting data, it can easily be incorporated into an analyses pipeline to predict the probability of a peptide occurring in a specific charge state of a peptide spectrum match, giving additional credibility towards the identifications. Additionally, CPred can be a relevant tool to look for peptides in the MS^1-layer, as it is possible to accurately predict a peptide's most abundant charge states, given its sequence and modifications.

Future development of the model could be focused on providing additional training data of different proteases given the bias towards trypsin. By doing so, CPred's flexibility will increase tremendously. A second scope of investigation may be extending the feature space by adding structural information of peptides, both for the raw sequence and the modifications. The structure of a peptide may play a pivotal role in the possibility of binding a proton. Interesting future research would also be looking variation of the charge state distribution of peptides, both inter- and intra-laboratory. A potential dataset for this is the benchmark dataset of Van Puyvelde et al. [36]. A final scope for future research is the extension of CPred to other biomolecules such as glycans, metabolites or other small molecules with their own set of features. The choice for developing CPred for peptides first is because of the homogeneity in bottoms-up proteomics workflows, leading to stable charge state distributions, and the availability of large datasets. Retraining CPred for other molecules comes with the prerequisite that we homogenise the analytical set-ups and have sufficiently large datasets available.

CPred is freely available online as a Python package with command-line possibilities under an Apache-2.0 software license. All code used within this research together with a developer version of the software is stored on GitHub (https://github.com/VilenneFrederique/CPred).

Data availability

All data supporting this study's findings are publicly available online.

Supporting Information

Supporting Information 1 provides supplementary details regarding the experimental procedures, including a comprehensive table listing all features in the model and a detailed technical description of the model's architecture and development process. Furthermore, it includes additional figures illustrating the outcomes of data exploration and model evaluation, thereby complementing the results presented in the main manuscript.

Acknowledgements

This research was funded by the Research Foundation – Flanders (FWO) under the "Beyond the Genome: Ethical Aspects of Large Cohort Studies" project (Case number G070722N). The resources and services used in this work were provided by the VSC (Flemish Supercomputer Centre), funded by the Research Foundation – Flanders (FWO) and the Flemish Government.

Conflicts of interest

The authors have declared no conflicts of interest.

Formatted: English (United States)

References

[1] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64-71. DOI: https://doi.org/10.1126/science.2675315.

[2] Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T.; Matsuo, T. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **1988**, *2* (8). DOI: https://doi.org/10.1002/rcm.1290020802.

[3] Ho, C. S.; Lam, C. W. K.; Chan, M. H. M.; Cheung, R. C. K.; Law, L. K.; Lit, L. C. W.; Ng, K. F.; Suen, M. W. M.; Tai, H. L. Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. *Clin. Biochem. Rev.* **2003**, *24* (1), 3-12.

[4] Downward, K. M.; Biemann, K. The Effect of Charge State and the Localization of Charge on the Collision-induced Dissociation of Peptide Ions. *Journal of the American Society for Mass Spectrometry* **1994**, *5* (11), 966-975. DOI: https://doi.org/10.1016/1044-0305(94)80015-4.

[5] Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399-1406. DOI: 10.1002/1096-9888(200012)35:12<1399::Aid-jms86>3.0.Co;2-r From NLM.

[6] Tsaprailis, G.; Nair, H.; Somogyi, Á.; Wysocki, V. H.; Zhong, W.; Futrell, J. H.; Summerfield, S. G.; Gaskell, S. J. Influence of Secondary Structure on the Fragmentation of Protonated Peptides. *J. Am. Chem. Soc.* **1999**, *121* (22), 5142-5154. DOI: 10.1021/ja982980h.

[7] Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **2007**, *6* (11), 1942-1951. DOI: 10.1074/mcp.M700073-MCP200 From NLM.

[8] Covey, T. R.; Bonner, R. F.; Shushan, B. I.; Henion, J. The determination of protein, oligonucleotide and peptide molecular weights by ion-spray mass spectrometry. *RCM* **1988**, *2* (11), 249-256. DOI: 10.1002/rcm.1290021111 From NLM.

[9] Smith, R. D.; Loo, J. A.; Loo, R. R. O.; Busman, M.; Udseth, H. R. Principles and practice of electrospray ionization—mass spectrometry for large polypeptides and proteins. *Mass Spectrometry Reviews* **1991**, *10* (5), 359-452. DOI: https://doi.org/10.1002/mas.1280100504 (acccessed 2024/07/02).

[10] Guevremont, R.; Siu, K. W. M.; Le Blanc, J. C. Y.; Berman, S. S. Are the electrospray mass spectra of proteins related to their aqueous solution chemistry? *Journal of the American Society for Mass Spectrometry* **1992**, *3* (3), 216-224. DOI: 10.1016/1044-0305(92)87005-J.

[11] Kelly, M. A.; Vestling, M. M.; Fenselau, C. C.; Smith, P. B. Electrospray analysis of proteins: A comparison of positive-ion and negative-ion mass spectra at high and low pH. *Org. Mass Spectrom.* **1992**, *27* (10), 1143-1147. DOI: https://doi.org/10.1002/oms.1210271028 (acccessed 2024/07/02).

[12] Schnier, P. D.; Gross, D. S.; Williams, E. R. On the maximum charge state and proton transfer reactivity of peptide and protein ions formed by electrospray ionization. *Journal of the American Society for Mass Spectrometry* **1995**, *6* (11), 1086-1097. DOI: 10.1016/1044-0305(95)00532-3.

[13] Sadygov, R. G.; Hao, Z.; Huhmer, A. Charger: Combination of Signal Processing and Statistical Learning Algorithms for Precursor Charge-State Determination from Electron-Transfer Dissociation Spectra. *Analytical Chemistry* **2008**, *80* (2), 376-386. DOI: https://doi.org/10.1021/ac071332q.

[14] Carvalho, P. C.; Cociorva, D.; Wong, C. C. L.; Da Gloria da C. Carvalho, M.; Barbosa, V. C.; Yates, J. R. Charge Prediction Machine: Tool for Inferring Precursor Charge States of Electron Transfer Dissociation Tandem Mass Spectra. *Analytical Chemistry* **2009**, *81* (5), 1996-2003. DOI: 10.1021/ac8025288.

[15] Sharma, V.; Eng, J. K.; Feldman, S.; Von Haller, P. D.; MacCoss, M. J.; Noble, W. S. Precursor charge state prediction for electron transfer dissociation tandem mass spectra. *Journal of Proteome Research* **2010**, *9* (10), 5438-5444. DOI: 10.1021/pr1006685.

[16] Liu, H.; Zhang, J.; Sun, H.; Xu, C.; Zhu, Y.; Xie, H. The prediction of peptide charge states for electrospray ionization in mass spectrometry. In *Procedia Environmental Sciences*, 2011; Elsevier B.V.: Vol. 8, pp 483-491. DOI: 10.1016/j.proenv.2011.10.076.

[17] Guan, S.; Moran, M. F.; Ma, B. Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular and Cellular Proteomics* **2019**, *18* (10), 2099-2107. DOI: 10.1074/mcp.TIR119.001412.
[18] Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4* (6), 1534-1536. DOI: 10.1002/pmic.200300744 From NLM Medline.
[19] Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; et al. Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods* **2017**, *14* (3), 259-262. DOI: 10.1038/nmeth.4153.
[20] Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **2016**, *11* (12), 2301-2319. DOI: 10.1038/nprot.2016.136.
[21] Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: The proteomics identifications database. *PROTEOMICS* **2005**, *5*, 3537-3545. DOI: DOI 10.1002/pmic.200401303.
[22] Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; et al. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **2017**, *4* (6), 587-599. DOI: https://doi.org/10.1016/j.cels.2017.05.009.
[23] Davis, S.; Charles, P. D.; He, L.; Mowlds, P.; Kessler, B. M.; Fischer, R. Expanding Proteome Coverage with CHarge Ordered Parallel IonaNalysis (CHOPIN) Combined with Broad Specificity Proteolysis. *Journal of Proteome Research* **2017**, *16*, 1288-1299. DOI: http://dx.doi.org/10.1021/acs.jproteome.6b00915.
[24] Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **2019**, *16* (6), 509-518. DOI: 10.1038/s41592-019-0426-7.
[25] Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace, 2009.
[26] Black, S. D.; Mould, D. R. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Analytical Biochemistry* **1991**, *193* (1), 72-82. DOI: https://doi.org/10.1016/0003-2697(91)90045-U.
[27] Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nature Methods* **2021**, *18* (11), 1363-1369. DOI: 10.1038/s41592-021-01301-5.
[28] Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. *arXiv* **2015**. DOI: https://doi.org/10.48550/arXiv.1603.04467.
[29] Hochreiter, S.; Schmidhuber, J. LONG SHORT-TERM MEMORY. *Neural Comput.* **1997**, *9* (8), 1735-1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.
[30] Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**. DOI: https://doi.org/10.48550/arXiv.1502.03167.
[31] Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *JMLR* **2018**, *18*, 1-52. DOI: https://doi.org/10.48550/arXiv.1603.06560.
[32] Shannon, C. E. A Mathematical Theory of Communication. *BSTJ* **1948**, *27* (3), 379-423. DOI: https://doi.org/10.1002%2Fj.1538-7305.1948.tb01338.x.
[33] Krusemark, C. J.; Frey, B. L.; Belshaw, P. J.; Smith, L. M. Modifying the Charge State Distribution of Proteins in Electrospray Ionization Mass Spectrometry by Chemical Derivatization. *Journal of the American Society for Mass Spectrometry* **2009**, *20* (9), 1617-1625. DOI: https://doi.org/10.1016%2Fj.jasms.2009.04.017.
[34] Zolg, D. P.; Wilhelm, M.; Schmidt, T.; Médard, G.; Zerweck, J.; Knaute, T.; Wenschuh, H.; Reimer, U.; Schnatbaum, K.; Kuster, B. ProteomeTools: Systematic Characterization of 21 Post-translational Protein Modifications by Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) Using

Formatted: Dutch (Netherlands)

Synthetic Peptides. *Molecular & Cellular Proteomics* **2018**, *17* (9), 1850-1863. DOI: https://doi.org/10.1074%2Fmcp.TIR118.000783.
[35] Thingholm, T. E.; Palmisano, G.; Kjeldsen, F.; Larsen, M. R. Undesirable Charge-Enhancement of Isobaric Tagged Phosphopeptides Leads to Reduced Identification Efficiency. *Journal of Proteome Research* **2010**, *9* (8), 4045-4052. DOI: https://doi.org/10.1021/pr100230q.
[36] Van Puyvelde, B.; Daled, S.; Willems, S.; Gabriels, R.; de Peredo, A. G.; Chaoui, K.; Mouton-Barbose, E.; Bouyssié, D.; Boonen, K.; Hughes, C. J.; et al. A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Sci. Data* **2022**, *9* (126). DOI: https://doi.org/10.1038/s41597-022-01216-6.

Table of Contents Graphic