Appendix

A Proteomics Terminology and Acronyms

A.1 Proteomics Terminology

- Retention Time (RT): The time taken by a peptide to pass through a column. This is dependent on different peptide features such as hydrophobicity and using a column helps to separate peptides before being analysed by a mass spectrometer.
- Fragment ions: An ion formed by fragmentation of a peptide in the mass spectrometer
- b and y ions: The B and Y ions for a given peptide represent the two halves formed by splitting the original peptide between various amino acids.
- Neutral Loss (NL): The loss of small molecules from peptide [92].
- MS/MS: Tandem Mass Spectrometry.
- Andromeda Score: A probabilistic score assigned to a spectrum by MaxQuant [93] to indicate the certainty of the identification. A higher score indicates higher confidence.
- Amino acid side chain: The organic R group that is unique to each amino acid.
- Peptide linkage: The chemical bond between two peptides.
- Indexed Retention Time (iRT): iRT is calculated by choosing two or more reference peptides and regressing a line between their retention times and is a unit-less quantity [94].
- Permuted phosphosites: Peptides with known phosphosites where we permute the site of phosporylation (at the same residue, but different position in the peptide). An example would be P[UNIMOD:21]EPTIDE -> PEP[UNIMOD:21]TIDE,where we have acquired spectra for both variants .This is not merely permutation on the sequence only, but rather a new input-target (i.e. sequence-spectrum) pair. Thus, these permuted phosphosites are helpful to evaluate tools for modification localization, specially since phosporylation is one of the most common PTMs and fairly complex to localize properly.
- Precursors: Peptide/Charge combinations since one peptide can occur in multiple charges.
- Tryptic peptides: Proteins digested by trypsin enzyme (most commonly used) are tryptic peptides. These peptides show a distinct pattern as they always end with K or R amino acids.
- Non-Tryptic peptides: They are peptides generated by enzymes other than trypsin. These can also be peptides which are picked with different methods other than digestion such as HLA peptides [95].
- Peptide-Spectrum Match (PSM): A peptide-spectrum match (PSM) is the assignment of a specific peptide sequence to a tandem mass spectrum (MS/MS), which contains information about fragmented peptide sequences. There are various protocols for analyzing MS/MS, but the primary goal is to assign a single peptide sequence to each MS/MS spectrum in a dataset.

A.2 Post-Translational Modifications (PTMs)

Table 3 lists the various modifications used in the datasets together with the respective Unimod ID [30], name, abbreviation, residue, change in atomic composition and change in Mass

B Data Annotation

As discussed in the paper, we used an expert system to annotate the MS/MS spectra, which relies on domain-specific conditional rules. Table 4 lists the rules with their original name and number as described in the original expert system publication [54].

We applied the rules by following a sequential workflow: (1) annotate one spectrum at a time, (2) generate all possible fragment ions, (3) check for matches within the tolerance specified by the expert system. For neutral losses we annotate up to 2 consecutive neutral losses.

Table 3: PTMs abbreviations and Unimod IDs.

UnimodID	Residue	Abbreviation	PTM Name	Delta Atom Count	Monoistopic Mass (da)
1	N-term/K	(ac)	Acetylation	H_2C_2O	42.01
3	K	(bi)	Biotinylation	$H_{14}C_{10}N_2O_2S$	226.08
4	C	(cam)	Carbamidomethylation	H_3C_2NO	57.02
7	R	(cit)	Citrullination	$H_{-1}N_{-1}O$	0.98
21	S/T/Y	(ph)	Phosporylation	HO_3P	79.97
27	E	(py)	Pyro-glu from E	$H_{-2}O_{-1}$	-18.01
28	Q	(py)	Pyro-glu from Q	$H_{-3}N_{-1}$	-17.02
34	K/R	(me)	Methylation	H_2C	14.02
35	M	(ox)	Oxidation	O	15.99
35	P	(hy)	Hydroxylation	O	15.99
36	K/R	(dime)	Di-Methylation	H_4C_2	28.03
37	K	(tme)	Tri-Methylation	H_6C_3	42.05
43	S/T	(glc)	GlcNAC	$H_{15}C_8NO_6$	203.8
43	S/T	(gal)	GalNAC	$H_{15}C_8NO_6$	203.8
58	K	(pr)	Propionylation	H_4C_3O	56.03
64	K	(su)	Succinylation	$H_4C_4O_3$	100.02
121	K	(ubi)	Ubiquitinylation	$H_64C_4N_2O_2$	114.04
122	K	(fo)	Formylation	CO	27.99
354	Y	(ni)	Nitro	$H_{-1}NO_2$	44.99
737	N-term/K	(tm)	Tandem Mass Tag	$H_{20}C_{12}N_2O_2$	229.17
747	K	(ma)	Malonylation	$H_2C_3O_3$	86.00
1289	K	(bu)	Butyrylation	H_6C_4O	70.04
1363	K	(cr)	Crotonylation	H_4C_4O	68.03
1848	K	(gl)	Glutarylation	$H_6C_5O_3$	114.03
1849	K	(hy)	Hydroxyisobutyrylation	$H_6C_4O_2$	86.04

Table 4: Rules from the used expert system [54] for annotation of MS/MS spectra.

Rule	Name	Rule	Name
35	b-ion series	80	Neutral loss at M
36	y-ion series	81	Neutral loss at M(Ox)
44	Charge1+	82	Neutral loss at N
45	Charge2+	83	Neutral loss at Q
46	Charge3+	84	Neutral loss at R
49	Neutral loss at S(ph)	85	Neutral loss at S
50	Neutral loss at T(ph)	86	Neutral loss at T
74	Neutral loss at C	87	Neutral loss at V
75	Neutral loss at D	88	Neutral loss at W
76	Neutral loss at E	97	Priority B Rule
77	Neutral loss at I	98	Priority Y Rule
78	Neutral loss at K	105	Priority Neutral Loss Rule
79	Neutral loss at L		

We use multiple threads to annotate multiple raw files in parallel. We used an AMD EPYC 7452 processor with 50 cores. The total time for annotating the complete raw data of 3 TB is around 40 hours.

Our implementation of the annotation pipeline is available in a dedicated GitHub repository under the name Spectrum Fundamentals [55]. Utilities for reading and parsing the raw data are collected in a dedicated GitHub repository under the name Spectrum IO [91].

C Further Exploratory Data Analysis

Tables 5, 6, 5, 8 provide further statistics on the modifications that exist in each of the three datasets with PTMs. The reported counts represent the number of unique sequences in total and those with at least one modification. All peptide sequences in the TMT dataset have a TMT modification on the N-term and all occurences of lysine (K). This specific modification is excluded from the table.

Table 5: Summary statistics of TMT dataset.

Residue	Modification	Unique Peptides	Precursors	Spectra
N-Term	TMT	641 K	742 K	24 M
K	TMT	359 K	414 K	12.6 M

Table 6: Summary statistics of Multi-PTM dataset.

Residue	Modification	Unique Peptides	Precursors	Spectra
K	Acetylation	52.1 K	71.2 K	3.1 M
N-Term	Acetylation	6.2 K	7.3 K	323 K
R	Citrullination	3.2 K	5 K	558 K
K	Methylation	1.9 K	2.8 K	160 K
R	Methylation	14.8 K	21.2 K	1.3 M
S	OGalNAc	1.2 K	1.8 K	185 K
T	OGalNAc	2 K	3.2 K	301 K
S	OGlcNAc	465	770	191 K
T	OGlcNAc	331	554	151 K
S	Phosphorylation	33.1 K	37.8 K	1.5 M
T	Phosphorylation	14 K	16 K	454 K
Y	Phosphorylation	31 K	38 K	3.1 M
E	Pyro-glu	2.8 K	3.3 K	173 K
Q	Pyro-glu	6.9 K	8.1 K	358 K
K	Ubiquitinylation	76.2 K	125.6 K	3.1 M

Table 7: Summary statistics of TMT-PTM dataset.

Residue	Modification	Unique Peptides	Precursors	Spectra
K	Acetylation	25 K	27.2 K	693 K
R	Methylation	11.7 K	13.5 K	516 K
S	Phosphorylation	26.9 K	31.6 K	1.1 M
T	Phosphorylation	11.1 K	12.2 K	248 K
Y	Phosphorylation	29.2 K	36.1 K	3.3 M

Table 9 highlights interesting patterns observed in Figure 3. First, the same modification occurring at different residues can have varying effects on the peptide properties, implying that including amino acid PTM information is essential to achieve better predictions. Second, some modifications have the same Unimod ID and the same molecular structure but only differ in their stereo-chemistry (spatial arrangement of atoms), yet they impact the peptide properties differently. Such scenarios are present in modified sequences and require a proper representation of PTMs (via encoding and domain-specific features) to predict peptide properties accurately. Table 11 in Appendix Section D shows the impact of PTMs on retention time for the special cases from Table 9.

Table 8: Summary statistics of Test-PTM dataset.

Residue	Modification	Unique Peptides	Precursors	Spectra
K	Acetylation	198	237	47.2 K
K	Biotinylation	197	225	25.7 K
K	Butyrylation	200	241	47.5 K
K	Crotonylation	200	237	48.5 K
K	Di-Methylation	189	348	39.2 K
K	Formylation	197	229	50.3 K
K	Glutarylation	200	233	52.9 K
K	Ubiquitinylation	200	382	52.4 K
K	Hydroxyisobutyrylation	199	226	48.4 K
K	Malonylation	198	224	35.1 K
K	Methylation	194	365	45.2 K
K	Propionylation	200	236	56.5 K
K	Succinylation	197	233	46.9 K
K	Tri-Methylation	186	329	38.1 K
P	Hydroxylation	169	235	31.5 K
R	Citrullination	184	247	37.9 K
R	Dimethyl-asymmetric	181	313	38.3 K
R	Dimethyl-symmetric	177	301	34.1 K
R	Methylation	179	308	41.3 K
Y	Nitro	175	215	58.5 K
Y	Phosphorylation	174	217	101 K
N-Term	TMT	6.2 K	7.6 K	32 K
K	TMT and Ubiquitinylation	38.5 K	51.7 K	756 K

Table 9: Examples of special amino acid-PTM pairs in our datasets.

Modification (PTM)	Residue	Scenario
Phosphorylation	S / T / Y	Same PTM, different residue
Di-methylation	R / K	Same PTM, different residue
GlcNAC/GalNAC	S / T	Same PTM, different structure
Symmetric/Asymmetric di-methylation	R	Same PTM, different structure

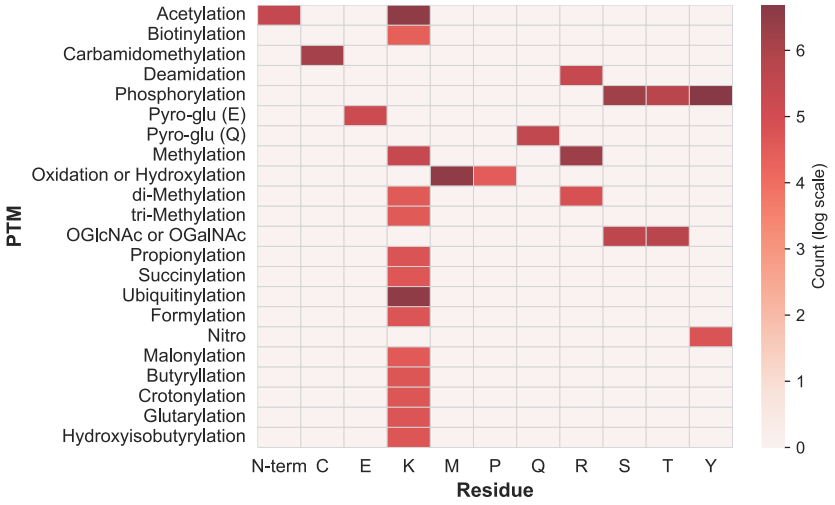

Figure 5: Heatmap indicating the frequency of each PTM occurring on different amino acids.

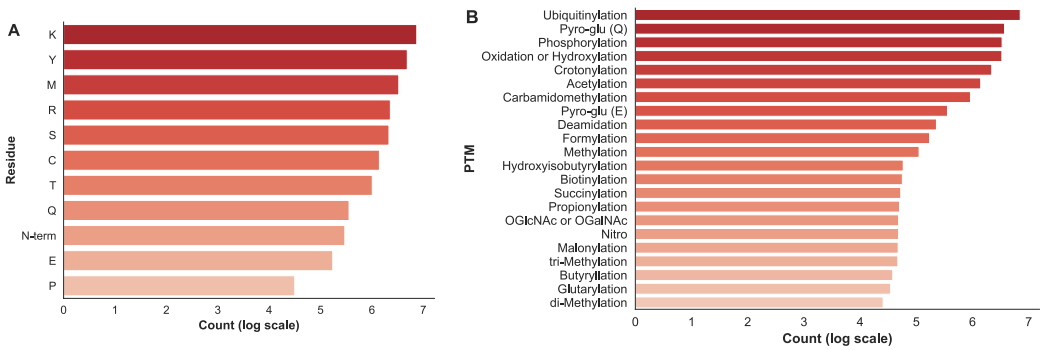

Figure 6: *A*: Frequency of each amino acid being reported as a modified site in the datasets. *B*: Frequency of occurrence of PTMs in the datasets.

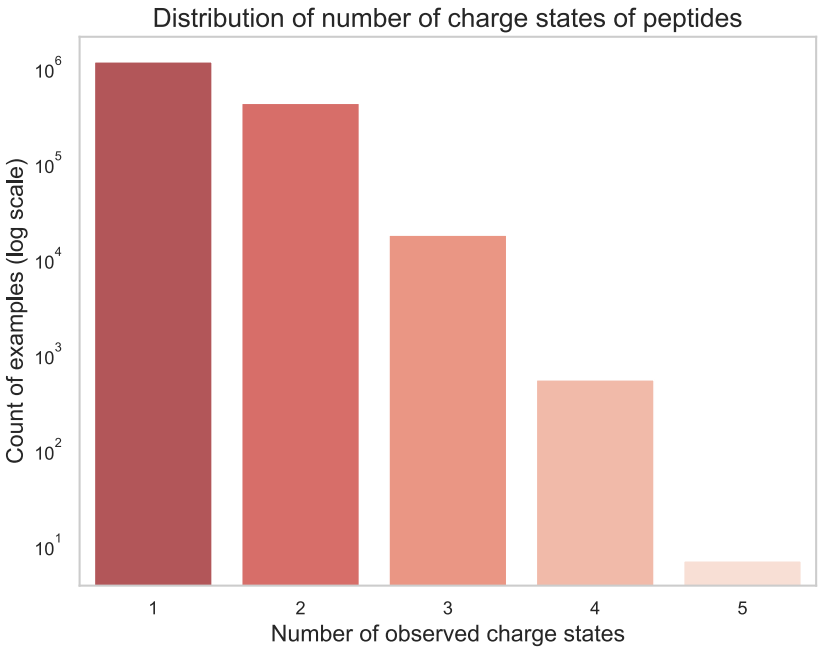

Figure 7: Bar plot of number of charge states for peptides in the datasets.

Table 10: Number of distinct charge states for peptides in the dataset.

Number of Charge States	Number of Examples
1	1,172,254
2	431,161
3	17,958
4	549
5	7

D Evaluation and Metrics

For retention time prediction, the time delta at 95% $\Delta t_{95\%}$ is the minimal time window containing the errors (residuals) between observed and predicted retention times for 95% of the peptides [78].

The 95% threshold corresponds to 2σ of the residuals. This threshold can be increased to a higher percentage for stricter evaluation of model performance [21].

For intensities, the Spectral Angle is defined as follows for V_a and V_b being the observed and predicted intensity vectors [21]:

$$SA = 1 - \frac{2}{\pi} \cos^{-1}\left(\frac{V_a \cdot V_b}{\|V_a\| \cdot \|V_b\|}\right)$$

We provide code to compute the metrics in our data GitHub repository [62].

D.1 Metrics for Datasets with PTMs

Throughout the paper, we calculated the Slope of a linear fit between two sets of iRT values. If the two sets of iRT values are perfectly aligned, we would expect a slope of 1 (e.g. model predictions against experimental values). We used this to highlight the impact of PTMs on retention time.

Although the two reported metrics (time delta 95 and Spectral angle) can be used for datasets with both unmodified and modified peptide sequences, some minor adaptations and granular evaluation can be conducted.

For example, when calculating the spectral angle, we align the peaks based on the annotation labels (y1, b1, y2, b2, etc...) to account for the m/z shift introduced by PTMs. Therefore, we only calculate how close the peak intensities are to each other.

For other tasks, reporting a suitable metric on subgroups of the dataset would give more insights into a model's performance. Potential subgroups include unmodified versus modified peptides, peptides with different modification types or peptides with modifications at different residues.

In Figure 8, we show iRT and intensity spectra behavior for different PTMs on TMT labeled peptides. We compare unmodified TMT-labeled peptides with modified TMT-labeled peptides. We also include the effect of TMT labeling on unmodified peptides. In Figure 9, we can see that PTMs with TMT show different behavior than PTMs on unlabeled peptides. Each point in the plot represents the effect of a single PTM, combined effects are not taken into account.

Tables 11, 12, 13, 14, 15, and 16 show the performance metrics reported on special cases and interesting subgroups of the datasets.

Table 11: Effect on retention time for the special residue-PTM pairs

Residue	Modification	Slope	iRT95	R2
S	Phosphorylation	0.85	21.1	0.83
T	Phosphorylation	0.88	20.2	0.87
Y	Phosphorylation	0.85	23.4	0.81
R	Di-Methylation	0.91	8.7	0.97
K	Di-Methylation	0.98	15.2	0.94
S	GlcNAC	0.92	15.8	0.9
S	GalNAC	1.04	13.2	0.9
T	GlcNAC	0.96	16.8	0.88
T	GalNAC	1.05	15.9	0.9
R	Symmetric Di-Methylation	0.91	8.7	0.97
R	aSymmetric Di-Methylation	0.93	11.0	0.97

D.2 Evaluation

Figure 10 shows the different distributions of the iRT residuals grouped by amino acid-PTM pair, sorted in ascending order by the delta iRT (difference between label and Prosit-DeltaAtoms prediction values for iRT). The results from Prosit model indicate that the model performs better than the DeepLC on most PTMs. Although there are some PTMs that both models struggle to predict, it shows that there is still room for improvement on the current SOTA models. An additional complexity can be observed on the Acetylated Lysine, showing a bi-modal distribution of delta iRT. Moreover, there is no consistent pattern among different PTMs, as they can shift the iRT to an earlier or a later point.

Table 12: Effect on retention time for the same PTM occurring at the same residue.

Residue	Modification	Slope	iRT95	R2
K	Acetylation	0.97	18.2	0.93
K	Ubiquitinylation	1.01	15.7	0.95
K	Methylation	1.1	18.8	0.94
R	Citrullination	0.85	16.2	0.97
R	Methylation	0.92	9.5	0.96
Y	Phosphorylation	0.84	24.2	0.8

Table 13: Effect on retention time for the same PTM occurring at a different residue.

Residue	Modification	Slope	iRT95	R2
P	Hydroxylation	0.94	9.9	0.97

Table 14: Effect on retention time for peptides with multiple PTMs

Number of different PTMs	Slope	iRT95	R2
1	0.93	18.7	0.95
2	0.89	20	0.95
3	0.88	20.5	0.94
4	0.86	22.6	0.92

Table 15: Effect on retention time for different PTMs

Residue	Modification	Slope	iRT95	R2
K	Biotinylation	1.19	38.1	0.94
K	Butyrylation	1.02	33.1	0.91
K	Crotonylation	1.04	34.4	0.91
K	Di-Methylation	0.98	15.2	0.94
K	Formylation	1.01	19	0.93
K	Glutarylation	1.07	21.5	0.93
K	Hydroxyisobutyrylation	0.96	24.9	0.93
K	Malonylation	0.94	20.7	0.93
K	Propionylation	0.99	23	0.92
K	Succinylation	1.05	22.7	0.94
K	Tri-Methylation	0.96	15.1	0.95
R	Dimethyl-asymmetric	0.93	11.1	0.97
R	Dimethyl-symmetric	0.91	8.7	0.97
Y	Nitro	0.91	29	0.84

Table 16: Effect on retention time for unmodified peptides.

Residue	Modification	Slope	iRT95	R2
K	unmodified	0.96	5.3	0.98
R	unmodified	0.95	4.4	0.97
Y	unmodified	0.96	6.2	0.97

Figure 11 shows the different distributions of the spectral angle grouped by amino acid-PTM pair, sorted in descending order by the median spectral angle with Prosit-DeltaAtoms predictions. A spectral angle below 0.6 usually indicates low similarity between the experimental and the predicted spectra, and hence, the predictions are not helpful in downstream tasks. Here, we divide the figure into three sub-plots to indicate which PTMs were used to train Prosit and which were used to train AlphapeptDeep. In Figure 11A, both models accurately predict the PTMs seen during

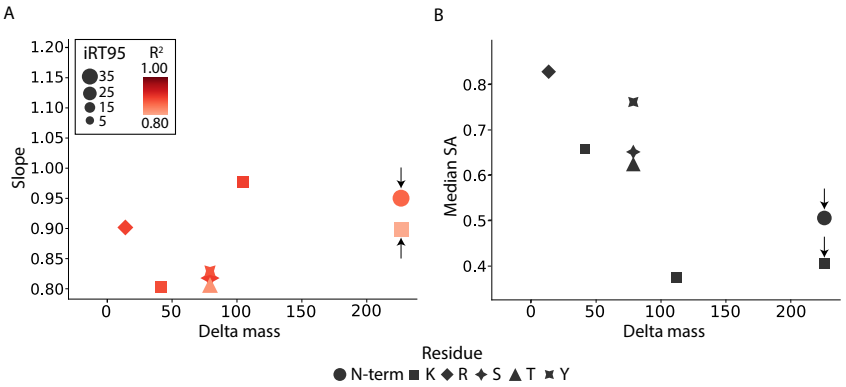

Figure 8: Scatter plots summarizing the iRT and fragment ion intensity difference between modified and matching unmodified TMT labeled peptides for all PTMs contained in the TMT dataset. Marked points show the difference between TMT labeled and unlabeled peptides.

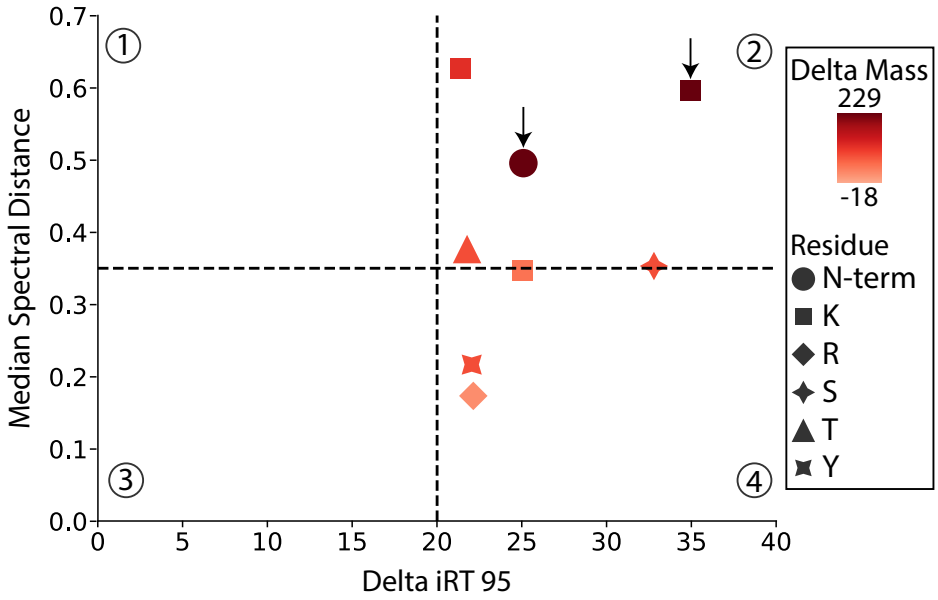

Figure 9: Impact of modifications on iRT vs. Intensity on TMT labeled peptides. Marked points show the difference between TMT labeled and unlabeled peptides.

training. In part B, AlphapeptDeep seems to be struggling with extrapolating on the unseen mods, while Prosit still performs well since these mods were included for the training phase. In Part C Prosit model generalizes to most of the unseen PTMs while still not getting the same performance as AlphapeptDeep. However, the model still struggles with some modifications: the first one is Biotinylation, and this might be because it has a bigger mass than most other PTMs and the model was not exposed to such modifications, the second is Malonylation, and this is mainly because, with this modification, there are very intense neutral loss peaks generated which changes the y- and b-ions intensity distribution entirely.

In Figure 12, Prosit shows comparable performance, while AlphapeptDeep seemingly struggles with this fragmentation method. The loss of performance in the AlphapeptDeep model is likely due to not having an input for the fragmentation method, and thus, it doesn't differentiate between HCD and CID fragmentation methods.

Figure 13 shows the different distributions of the precursor charge distributions grouped by amino acid-PTM pair, sorted in ascending order by the MAE (difference between label and Prosit prediction values for precursor charge distribution). We visualize how different modifications affect the charge

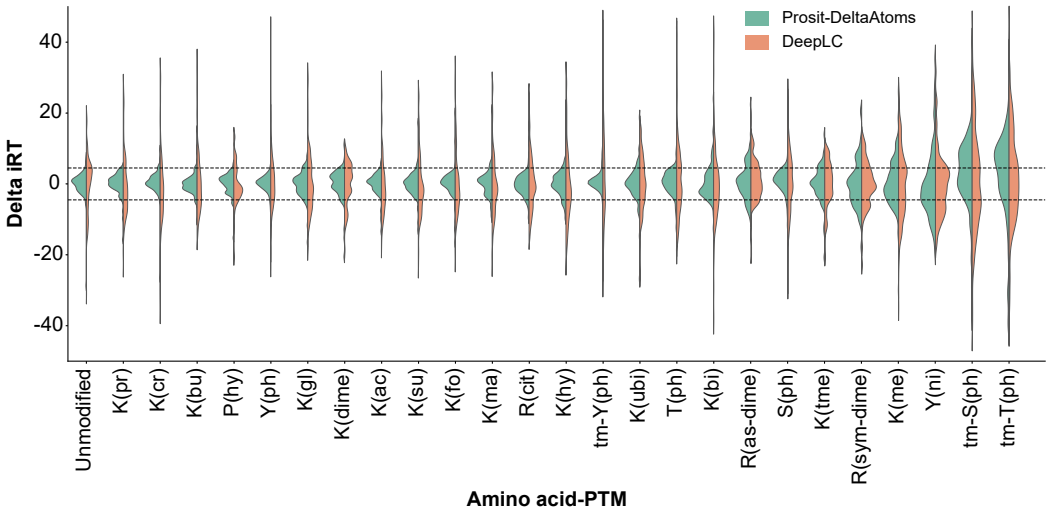

Figure 10: Retention Time prediction with Prosit and DeepLC violin plot with dashed lines indicating the delta iRT 95 for unmodified predictions with Prosit. Sequences are grouped by modifications.

distribution differently; some almost have no effect, and others have more prominent effects, leading to MAE values reaching 0.5.

E Elaboration on Supported Tasks

This dataset in combination with the original PROSPECT dataset [21] offers the opportunity to analyze in details the effect of PTMs on different peptide properties. While we covered in the main RT and fragment ion intensity, this can also be used for more such as precursor charge and neutral loss patterns [96, 97, 98]. It also provide the option of studying how the PTM location affects the peptide properties as well, while there are multiple algorithms trying to do this task [46, 47, 41], this is rather understudied with no proper study of the different effects of such change.

In Table 17, we provided a list of machine learning tasks that our datasets can primarily support in the context of PTMs. The datasets can be also used for several other tasks, either as-is or with the standard respective pre-processing required for the machine learning task at hand, without the need for deep domain knowledge or further annotations. Table 17 lists all tasksas feasible with our datasets to the best of our knowledge.

F Downstream Impact

Figure 14 shows that a model trained on PROSPECT PTM can lead to a gain of 20 % in PSMs and peptides after rescoring [108].

G More Details on Splitting the Datasets

Our general recommendation as described in Section 3.3 is to use Test-PTM as a hold-out dataset and split the three remaining datasets for training and validation of models. However, while iterating on model development and training, we additionally recommend splitting the three datasets into training, validation and test splits based on uniqueness of unmodified sequences (sequence-based disjoint split), where examples for the same unmodified peptide sequence should appear in only one of the splits [33]. Afterwards, the final selected model can be evaluated on the Test-PTM as a hold-out dataset.

For retention time, users of the dataset should filter out the quality control (QC) peptides that are used for RT calibration. This is detailed in [94].

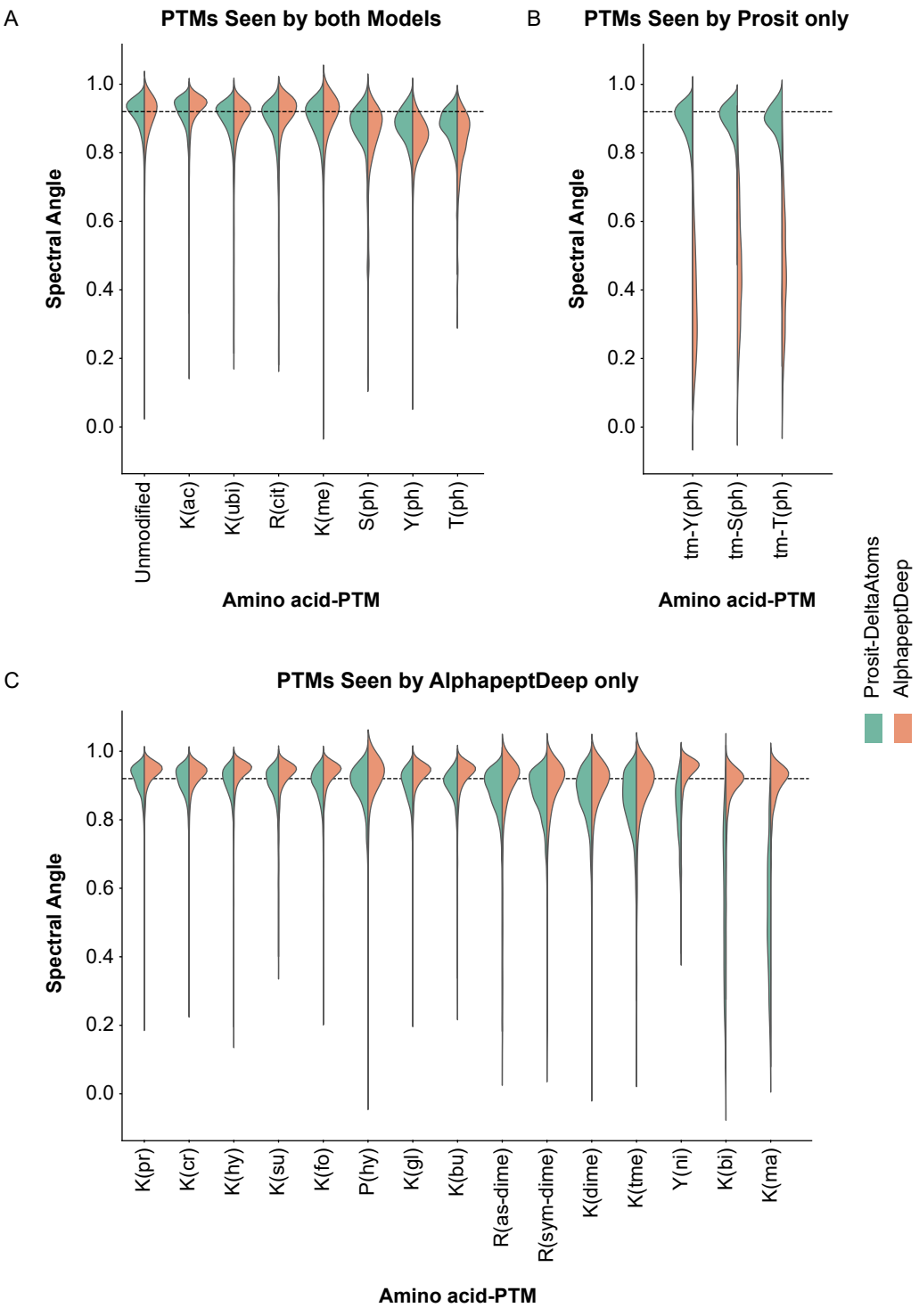

Figure 11: Violin plot with intensity predictions from Prosit and AlphaPeptDeep for HCD fragmentation. The dashed line indicating median SA for predictions on unmodified peptides with Prosit.

The counts in Table 18 show that the datasets contain mostly tryptic peptides. However, the TMT and the PTM datasets contain some non-tryptic peptides. Since studying PTMs in the context of tryptic peptides is more common, there should be no significant concerns about selection bias. Additionally,

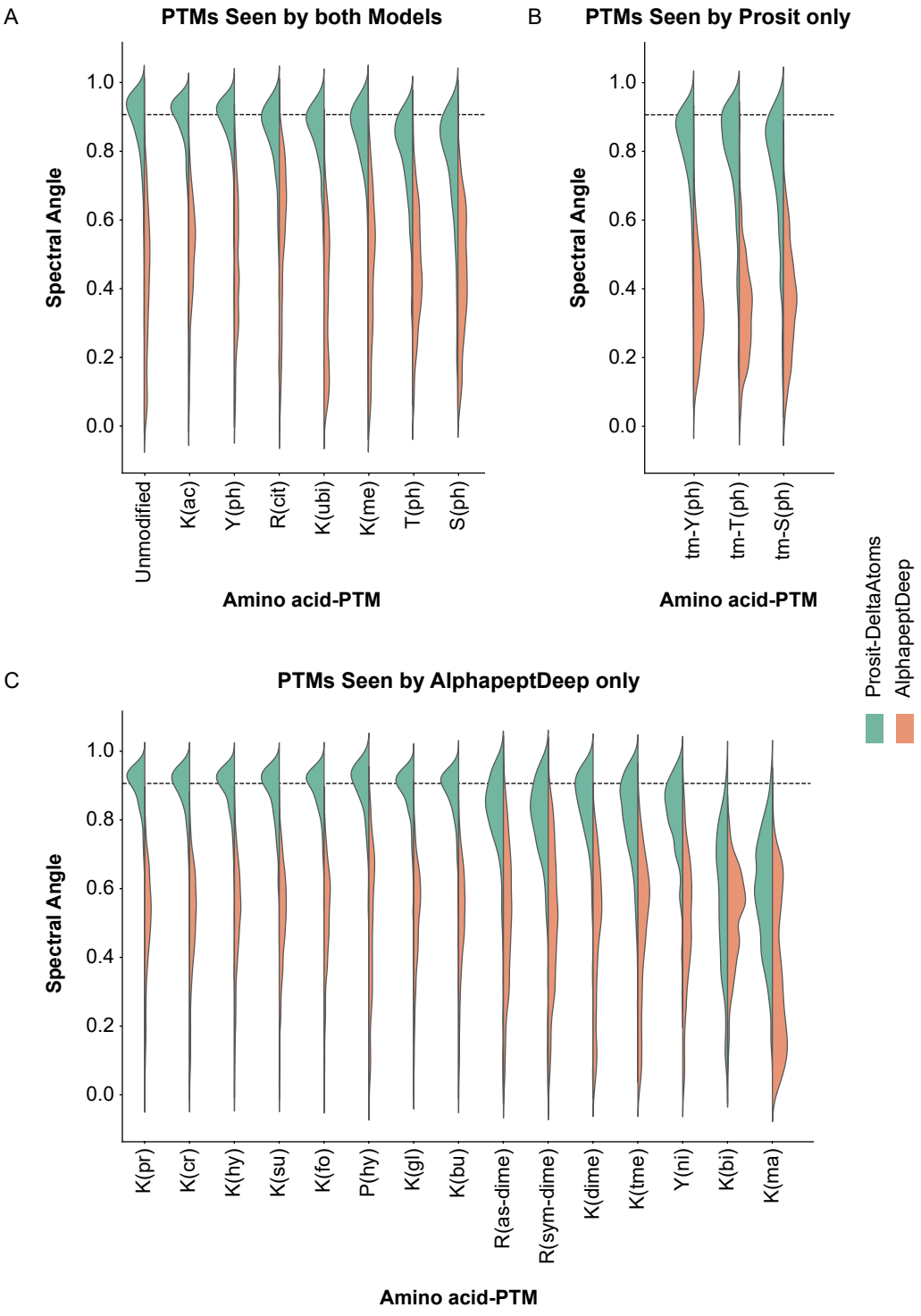

Figure 12: Violin plot with intensity predictions from Prosit and AlphaPeptDeep for CID fragmentation. The dashed line indicates the median SA for predictions on unmodified peptides with Prosit.

training models with the new datasets together with the unmodified peptides from PROSPECT [21] would help alleviate bias towards tryptic peptides.

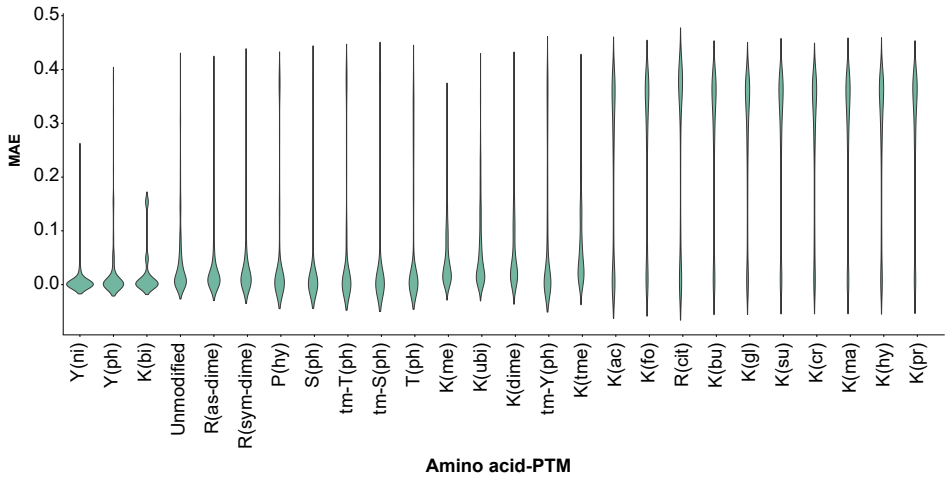

Figure 13: Distribution of predictions for precursor charge states.

Table 17: Listing of tasks feasible with PROSPECT PTM datasets.

Name	Input	Target	Type	References
PTM site prediction	Sequence	PTM Site	Classification	[99, 100, 53]
Retention Time	Sequence	RT	Regression (single value)	[33, 38, 51, 40]
Retention Time with PTMs	Sequence	RT	Regression (single value)	[37]
Retention Time	Sequence	RT	Regression (distribution)	-
Intensity prediction	Sequence	intensities	Regression (vector)	[33, 34, 101]
Intensity prediction with PTMs	Sequence	intensities	Regression (vector)	[41, 39]
Fragment Presence	Sequence	Present/Not	Binary classification	[102]
Charge prediction	Sequence	Charge distribution	Regression	[69, 70]
De novo sequencing	Intensity	Sequence	Classification Ranking	[86, 87, 103, 104]
Sequence/Spectral Embedding	Sequence Spectra	Embeddings	Representation Similarity learning	[105, 84, 106]
Multiple properties	Sequence	RT charge intensity	Multi-task Learning	[69]
Sequence Clustering	Sequence	Cluster	Unsupervised Clustering	[107]

H Experimental Details

For the evaluation of peptide sequences with PTMs, we trained various models. The first model is Prosit baseline [33]. The model was trained on the original PROSPECT dataset [21] for both

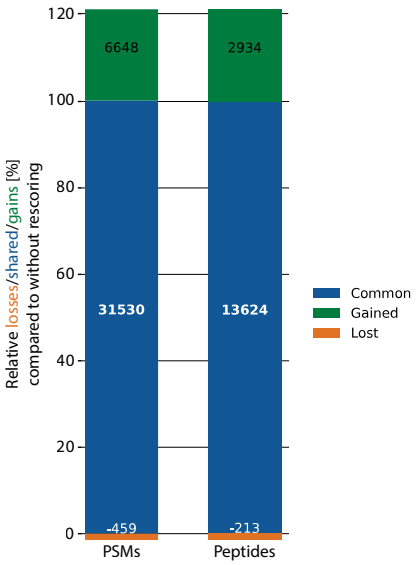

Figure 14: Stacked bar chart showing the number of confident PSMs (Peptide-Spectrum-Matches) (left) and peptides (right) below 1% FDR (False Discovery Rate) lost (orange), shared (blue) and gained (green) when rescoring using (Prosit) compared to the MSFragger [61] results.

Table 18: Count of Tryptic versus non-tryptic peptides.

Dataset	Tryptic Peptides	Non-Tryptic Peptides	Tryptic Spectra	Non-Tryptic Spectra
TMT	396 K	318 K	19 M	9.2 M
PTM	300 K	7.2 K	19.4 M	267 K
TMT-PTM	157 K	1.6 K	7.7 M	52 K

tasks, retention time prediction, and intensity prediction; this model serves as a baseline performance of how a model would predict different features without PTM encoding. Then, we trained other models using different ways of PTM encoding to see which might show the best performance. The Prosit naive model was trained on all datasets except Test-PTM. Hence, it does not support unseen PTMs but instead ignores them. Two variants of Prosit encoded PTMs with domain-specific features: Prosit-DeltaMass and Prosit-DeltaAtoms. Prosit-DeltaMass uses the mass introduced by the PTM as an input feature to the model, while Prosit-DeltaAtoms uses the atom count introduced by the PTM [40, 37]. AlphaPeptDeep [40] does not support N-term modifications.

The training and the inference were conducted using a single Nvidia A30 GPU. The Prosit model is an encoder-decoder RNN-based model that can predict indexed retention time and intensity spectrum for a given peptide sequence. The recurrent layers used in the architecture are Gated Recurrent Units (GRUs). More details on the architecture can be found in [33]. Training time is in the range of 2-3 hours on the unmodified dataset for retention time and 25 to 30 hours for MS/MS spectra.

For comparison, we ran inference on a DeepLC model for retention time prediction using an Nvidia A30 GPU. The DeepLC model architecture encodes input sequences and atom counts from amino acids and PTMs using 1D-convolutional and max pooling layers. The DeepLC model architecture includes a branch of fully-connected layers for global features engineered manually before training from the sequences [37]. Figure 15 depicts the difference between DeepLC predictions and the experimental iRT.

For MS spectrum prediction, we ran inference on AlphaPeptDeep [40] for MS spectra prediction using an Nvidia A30 GPU. The AlphaPeptDeep frameworks provides several model architectures. The pre-trained model we used encodes input sequences and atom counts from amino acids and PTMs using transformer layers with positional encoding.

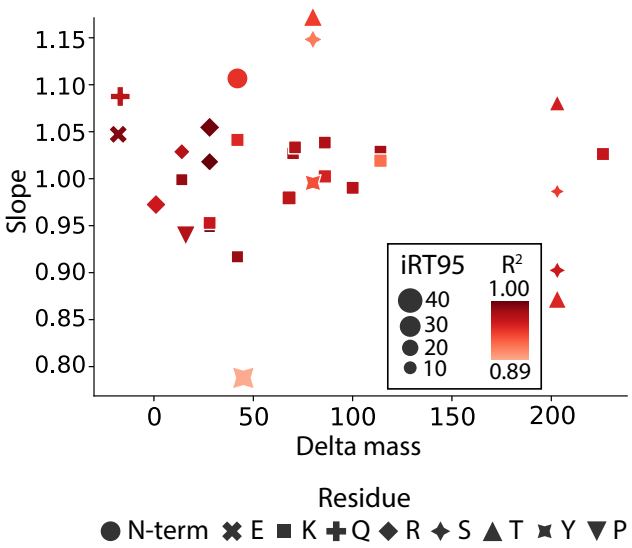

Figure 15: The iRT difference between DeepLC predictions on modified and matching unmodified peptides versus m/z values.

I Dataset Documentation: Datasheet for Datasets

I.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The purpose is to introduce and establish multiple reference datasets for peptide sequences with different groups of Post-Translational Modifications present, complementing the original PROSPECT dataset, which contains unmodified peptide sequences. As of June 2024, the original PROSPECT dataset [28] was viewed on Zenodo 871 times and downloaded 1116 times.

The four new datasets are processed and curated for machine learning in Proteomics. Although they are not constrained to specific tasks, the focus is on three common tasks in proteomics; retention time prediction, MS/MS spectrum prediction, and precursor charge state prediction.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Computational Mass Spectrometry Chair at the School of Life Sciences, Technical University of Munich, Germany.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation was partially funded by the following grants: European Proteomics Infrastructure Consortium providing access, Grant Number 823839 and Bundesministerium für Bildung und Forschung – BMBF, Grant Number 031L0008A.

Other comments?

We aim that this would be a start for different groups to release curated dataset with machine learning tasks in mind instead of only publishing raw datasets that required several processing steps before being useful for machine learning.

I.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people

and interactions between them; nodes and edges)? Please provide a description. How many instances are there in total (of each type, if appropriate)?

Here we publish 4 different datasets. Instances of each dataset are peptide sequences, their corresponding annotations, and meta-data for spectra. We have in 4.6B unique peaks for 58.6M spectra of 1.2M unique peptides with 30 unique PTM-residue combinations. We uploaded 2 different file types; one for meta data for each spectrum and another with annotations.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

We have a dataset of all valid identifications from ProteomeTools PTMs and TMT raw data [48], one of the largest datasets with synthetic peptides.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

We have the unprocessed meta data that we get from raw files generated as output from the mass spectrometer. We process the spectra from MS with the identifications that we get from MaxQuant [93] to annotate our dataset and annotate the fragment ions found in the spectra.

Is there a label or target associated with each instance? If so, please provide a description.

There are various machine learning problem formulations in proteomics, more details are in section 2. For the three tasks we focused on; namely, retention time, precursor charge state, and intensity prediction, the targets are retention time (and indexed retention time), precursor charge states, and the annotated spectra, respectively.

Instances of the dataset are generally linked together with the raw file and scan number associated with each spectra.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no direct relationships between different instances they might have some features in the metadata such as length, retention time, collision energy, PTM-residue combination and peptide sequence.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

For MS Spectra, we recommend splitting data based on the peptide sequence so no peptide sequence is shared across different splits to avoid data leakage. Also splitting each pool in different files as each pool has different set of peptides to ensure that the splits has all the different types of peptides and didn't miss any. For charge state prediction, we recommend to split the data with a similar approach.

For retention time, while we suggest the same as Spectra in terms of not sharing sequences in different splits, we additionally recommend to filter examples and keep only one copy of each with the mean retention time of measurements for the same sequence. The mean retention time for each sequence can then be used for training the model.

The splits we used are available as ready-to-use datasets on the Hugging Face hub for the three tasks; retention time [63], MS2 [64], and charge state [65].

We suggest as well to use the PTM-Test as another holdout dataset for models trying to predict features for peptides with PTMs. We explain in details why this is useful in 3.3. More details are in Appendix Section G. The respective curated dataset on the Hugging Face hub is available in the three task repositories under a separate configuration denoted as *holdout*.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Our objective was to reduce the number of miss-identifications in the dataset, since we know which set of peptides we expect in each raw file, we remove all other identifications made by MaxQuant [93]. Although there might still be miss-identifications after this filtering, they would rather be less than 1%, which is the known acceptable cut off-in the field. There are redundancies as the same peptide would be measured multiple times but the spectra and Retention time would be slightly different in different measurements, we kept in this case all instances in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained as all the processed information is in one place. The only external resource is when users want to get access to the raw unprocessed data this is shared on pride archives [27, 25, 26]. All the archives are open access with a CC license and no restrictions on getting the data.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No

I.3 Collection

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Raw files were acquired with a Mass spectrometer and peptide identifications were made with MaxQuant [93] (a software for database search). Here we depend on MaxQuant for identifications, but as mentioned in the previous section we remove miss-identifications to decrease the number of wrong labels. This is a particular strength of this dataset since we know which peptides exist per sample, based on the fact that they were specifically synthesized. For other datasets, we might only know to which organism they belong, which can lead to a higher number of misidentifications.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

Data was measured with Mass spectrometers and annotated with a software. Already explained in the question above how the dataset was validated.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

No sampling was done, we include all peptides that were measured.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Raw data were acquired by PhD students working on the ProteomeTools project [48] and annotated and curated by PhD students and the authors of the accompanying paper.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Data acquisition started as early as 2017, but the time-frame doesn't affect the data in any shape or form.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, the data is based on ProteomeTools which contains only synthetic peptide samples.

I.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

No.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is publicly available through the PRIDE archives [27, 25, 26]. The annotated datasets we provide are available on Zenodo [56, 57, 58, 59]. The processed and split task-specific datasets are available on the Hugging Face Hub [63, 64, 65].

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

No, we used MaxQuant to remove misidentifications. Further processing of the datasets was done with Python scripts that are available in a dedicated dataset utilities GitHub repository [62].

I.5 Usage

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, parts of the dataset were previously used in different models for predicting fragment ions intensity and retention time, examples include the work in [43].

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

We referenced previous work in the paper and in Zenodo along with the dataset itself [57, 56, 58, 59].

What (other) tasks could the dataset be used for?

The data can be used for various tasks, examples include prediction of different peptide features, study double annotations for different peaks, assignment of annotations to peaks and localizing PTMs. For more details, please refer to Section 3 and Appendix Section E.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

No, not as far as we know.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No, not as far as we know.

I.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Both the raw data from ProteomeTools and our dataset PROSPECT are publicly available. Every third party outside the entity on behalf of which the dataset was generated has access to it now.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

We uploaded the 4 datasets to Zenodo, each with a dedicated DOI [57, 56, 58, 59]. The processed and split task-specific datasets are uploaded to the Hugging Face Hub with dedicated DOIs [63, 64, 65]. We also have a GitHub repository with utilities to download the dataset [62].

When will the dataset be distributed?

We published the datasets on Zenodo in October 2023 and June 2024. We published the accessible splitted, aggregated, and processed datasets to the Hugging Face Hub in June 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Open access, Creative Commons Attributions 4.0 International.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

I.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

Professorship for Computational Mass Spectrometry at the Technical University of Munich (TUM). The datasets are hosted on Zenodo [56, 57, 58, 59] and the task-specific processed datasets are hosted on the Hugging Face [63, 64, 65]. Current maintainers are the authors and later other members of the Professorship at TUM.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Mathias Wilhelm (mathias.wilhelm@tum.de).

Is there an erratum? If so, please provide a link or other access point. No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

If we detect further misidentifications or improve the quality of annotations, we will release subsequent versions with the respective updates. This will be versioned and announced in Zenodo, the Hugging Face Hub, and GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, the dataset is neither related to people nor based on human samples.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Yes, different versions will be maintained on Zenodo and the Hugging Face Hub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

We welcome and encourage others to extend/augment/build on/contribute to the dataset. We suggested initiating contact with our professorship and we will discuss the best options for communicating/distribution the additions.

J Author Statement

The authors confirm all responsibility in case of violation of rights and confirm the licence associated with the dataset.

References

[1] Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nature reviews Molecular cell biology*, 5(9):699–711, 2004.

[2] Steven R Shuken. An introduction to mass spectrometry-based proteomics. *Journal of Proteome Research*, 2023.

[3] Ting Huang, Jingjing Wang, Weichuan Yu, and Zengyou He. Protein inference: a review. *Briefings in bioinformatics*, 13(5):586–614, 2012.

[4] Marcus Bantscheff, Simone Lemeer, Mikhail M Savitski, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry*, 404(4):939–965, 2012.

[5] Oliver Pagel, Stefan Loroch, Albert Sickmann, and René P Zahedi. Current strategies and findings in clinically relevant post-translational modification-specific proteomics. *Expert review of proteomics*, 12(3):235–253, 2015.

[6] Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. *Database*, 2021, 2021.

[7] Mao Peng, Arjen Scholten, Albert JR Heck, and Bas van Breukelen. Identification of enriched ptm crosstalk motifs from large-scale experimental data sets. *Journal of proteome research*, 13(1):249–259, 2014.

[8] A Saskia Venne, Laxmikanth Kollipara, and René P Zahedi. The next level of complexity: crosstalk of posttranslational modifications. *Proteomics*, 14(4-5):513–524, 2014.

[9] Tony Hunter. The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Molecular cell*, 28(5):730–738, 2007.

[10] Matthew P Torres, Henry Dewhurst, and Niveda Sundararaman. Proteome-wide structural analysis of ptm hotspots reveals regulatory elements predicted to impact biological function and disease. *Molecular And Cellular Proteomics*, 15(11):3513–3528, 2016.

[11] Jana Zecha, Florian P Bayer, Svenja Wiechmann, Julia Woortman, Nicola Berner, Julian Müller, Annika Schneider, Karl Kramer, Mar Abril-Gil, Thomas Hopf, et al. Decrypting drug actions and protein modifications by dose-and time-resolved proteomics. *Science*, 380(6640):93–101, 2023.

[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex Bridgland, et al. Alphafold 2. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction; DeepMind: London, UK*, 2020.

[13] Federica Del Monte and Giulio Agnetti. Protein post-translational modifications and misfolding: New concepts in heart failure. *PROTEOMICS–Clinical Applications*, 8(7-8):534–542, 2014.

[14] Min-Gang Su, Julia Tzu-Ya Weng, Justin Bo-Kai Hsu, Kai-Yao Huang, Yu-Hsiang Chi, and Tzong-Yi Lee. Investigation and identification of functional post-translational modification sites associated with drug binding and protein-protein interactions. *BMC systems biology*, 11:69–80, 2017.

[15] Albert B Arul and Renã AS Robinson. Sample multiplexing strategies in quantitative proteomics. *Analytical chemistry*, 91(1):178–189, 2018.

[16] Jiaming Li, Zhenying Cai, Ryan D Bomgarden, Ian Pike, Karsten Kuhn, John C Rogers, Thomas M Roberts, Steven P Gygi, and Joao A Paulo. Tmtpro-18plex: the expanded and complete set of tmtpro reagents for sample multiplexing. *Journal of proteome research*, 20(5):2964–2972, 2021.

[17] Bo Wen, Wen-Feng Zeng, Yuxing Liao, Zhiao Shi, Sara R Savage, Wen Jiang, and Bing Zhang. Deep learning in proteomics. *Proteomics*, 20(21-22):1900335, 2020.

[18] Alexander Kensert, Gert Desmet, and Deirdre Cabooter. Molgraph: a python package for the implementation of small molecular graphs and graph neural networks with tensorflow and keras. *arXiv preprint arXiv:2208.09944*, 2022.

[19] Jun Xue, Bingyi Wang, Hongchao Ji, and WeiHua Li. Rt-transformer: Retention time prediction for metabolite annotation to assist in metabolite identification. *Bioinformatics*, page btae084, 2024.

[20] Qiyue Kang, Pengfei Fang, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. Deep graph convolutional network for small-molecule retention time prediction. *Journal of Chromatography A*, 1711:464439, 2023.

[21] Omar Shouman, Wassim Gabriel, Victor-George Giurcoiu, Vitor Sternlicht, and Mathias Wilhelm. PROSPECT: Labeled tandem mass spectrometry dataset for machine learning in proteomics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32882–32896. Curran Associates, Inc., 2022.

[22] Konstantin Weißenow, Michael Heinzinger, and Burkhard Rost. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*, 30(8):1169–1177, 2022.

[23] Zilong Hou, Yuning Yang, Zhiqiang Ma, Ka-chun Wong, and Xiangtao Li. Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning. *Communications Biology*, 6(1):73, 2023.

[24] Suresh Pokharel, Pawel Pratyush, Michael Heinzinger, Robert H Newman, and Dukka B Kc. Improving protein succinylation sites prediction using embeddings from protein language model. *Scientific Reports*, 12(1):16933, 2022.

[25] Daniel P Zolg and Kuster Bernhard. ProteomeTools. https://www.ebi.ac.uk/pride/archive/projects/PXD023119, 2021. [Online; accessed 31-May-2023].

[26] Daniel P Zolg and Kuster Bernhard. ProteomeTools. https://www.ebi.ac.uk/pride/archive/projects/PXD023120, 2021. [Online; accessed 31-May-2023].

[27] Daniel P Zolg and Kuster Bernhard. ProteomeTools. https://www.ebi.ac.uk/pride/archive/projects/PXD009449, 2018. [Online; accessed 31-May-2023].

[28] Omar Shouman, Wassim Gabriel, and Mathias Wilhelm. PROSPECT Dataset. DOI:https://doi.org/10.5281/zenodo.6602020, 2022.

[29] H BIELKA GDR, N Sharon, and EC WEBB Australia. Nomenclature and symbolism for amino acids and peptides. *Pure and Applied Chemistry*, 56:595–624, 1984.

[30] David M Creasy and John S Cottrell. Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4(6):1534–1536, 2004.

[31] Luisa Montecchi-Palazzi, Ron Beavis, Pierre-Alain Binz, Robert J Chalkley, John Cottrell, David Creasy, Jim Shofstahl, Sean L Seymour, and John S Garavelli. The psi-mod community standard for representation of protein modification data. *Nature biotechnology*, 26(8):864–866, 2008.

[32] John S Garavelli. The resid database of protein modifications as a resource and annotation tool. *Proteomics*, 4(6):1527–1533, 2004.

[33] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 16(6):509–518, 2019.

[34] Markus Ekvall, Patrick Truong, Wassim Gabriel, Mathias Wilhelm, and Lukas Käll. Prosit transformer: A transformer for prediction of ms2 spectrum intensities. *Journal of Proteome Research*, 2022.

[35] Anton A Goloborodko, Lev I Levitsky, Mark V Ivanov, and Mikhail V Gorshkov. Pyteomics—a python framework for exploratory data analysis and rapid software prototyping in proteomics. *Journal of The American Society for Mass Spectrometry*, 24(2):301–304, 2013.

[36] Lev I Levitsky, Joshua A Klein, Mark V Ivanov, and Mikhail V Gorshkov. Pyteomics 4.0: five years of development of a python proteomics framework. *Journal of proteome research*, 18(2):709–714, 2018.

[37] Robbin Bouwmeester, Ralf Gabriels, Niels Hulstaert, Lennart Martens, and Sven Degroeve. Deeplc can predict retention times for peptides that carry as-yet unseen modifications. *Nature methods*, 18(11):1363–1369, 2021.

[38] Wen-Feng Zeng, Xie-Xuan Zhou, Wen-Jing Zhou, Hao Chi, Jianfeng Zhan, and Si-Min He. Ms/ms spectrum prediction for modified peptides using pdeep2 trained by transfer learning. *Analytical chemistry*, 91(15):9724–9731, 2019.

[39] Ching Tarn and Wen-Feng Zeng. pdeep3: toward more accurate spectrum prediction with fast few-shot learning. *Analytical Chemistry*, 93(14):5815–5822, 2021.

[40] Wen-Feng Zeng, Xie-Xuan Zhou, Sander Willems, Constantin Ammar, Maria Wahle, Isabell Bludau, Eugenia Voytik, Maximillian T Strauss, and Matthias Mann. Alphapeptdeep: a modular deep learning framework to predict peptide properties for proteomics. *Nature Communications*, 13(1):7238, 2022.

[41] Yu Zong, Yuxin Wang, Yi Yang, Dan Zhao, Xiaoqing Wang, Chengpin Shen, and Liang Qiao. DeepFLR facilitates false localization rate control in phosphoproteomics. *Nature Communications*, 14(1):2269, 2023.

[42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[43] Wassim Gabriel, Matthew The, Daniel P Zolg, Florian P Bayer, Omar Shouman, Ludwig Lautenbacher, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, et al. Prosit-TMT: deep learning boosts identification of TMT-labeled peptides. *Analytical Chemistry*, 94(20):7181–7190, 2022.

[44] Kenneth Verheggen, Helge Ræder, Frode S Berven, Lennart Martens, Harald Barsnes, and Marc Vaudel. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. *Mass spectrometry reviews*, 39(3):292–306, 2020.

[45] Hong Li, Xiaobin Xing, Guohui Ding, Qingrun Li, Chuan Wang, Lu Xie, Rong Zeng, and Yixue Li. Sysptm: a systematic resource for proteomic research on post-translational modifications. *Molecular & Cellular Proteomics*, 8(8):1839–1849, 2009.

[46] David D Shteynberg, Eric W Deutsch, David S Campbell, Michael R Hoopmann, Ulrike Kusebauch, Dave Lee, Luis Mendoza, Mukul K Midha, Zhi Sun, Anthony D Whetton, et al. Ptmprophet: Fast and accurate mass modification localization for the trans-proteomic pipeline. *Journal of proteome research*, 18(12):4262–4272, 2019.

[47] Thammakorn Saethang, D Michael Payne, Yingyos Avihingsanon, and Trairak Pisitkun. A machine learning strategy for predicting localization of post-translational modification sites in protein-protein interacting regions. *BMC bioinformatics*, 17:1–15, 2016.

[48] Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, et al. Building proteometools based on a complete synthetic human proteome. *Nature methods*, 14(3):259–262, 2017.

[49] Daniel Paul Zolg, Mathias Wilhelm, Tobias Schmidt, Guillaume Médard, Johannes Zerweck, Tobias Knaute, Holger Wenschuh, Ulf Reimer, Karsten Schnatbaum, and Bernhard Kuster. Proteometools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (lc-ms/ms) using synthetic peptides. *Molecular & Cellular Proteomics*, 17(9):1850–1863, 2018.

[50] Kaiyuan Liu, Sujun Li, Lei Wang, Yuzhen Ye, and Haixu Tang. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. *Analytical Chemistry*, 92(6):4275–4283, 2020.

[51] Chunwei Ma, Yan Ren, Jiarui Yang, Zhe Ren, Huanming Yang, and Siqi Liu. Improved peptide retention time prediction in liquid chromatography through deep learning. *Analytical chemistry*, 90(18):10881–10888, 2018.

[52] Daniel J Geiszler, Daniel A Polasky, Fengchao Yu, and Alexey I Nesvizhskii. Detecting diagnostic features in ms/ms spectra of post-translationally modified peptides. *Nature Communications*, 14(1):4132, 2023.

[53] Brandon M Gassaway, Jiaming Li, Ramin Rad, Julian Mintseris, Kyle Mohler, Tyler Levy, Mike Aguiar, Sean A Beausoleil, Joao A Paulo, Jesse Rinehart, et al. A multi-purpose, regenerable, proteome-scale, human phosphoserine resource for phosphoproteomics. *Nature Methods*, pages 1–5, 2022.

[54] Nadin Neuhauser, Annette Michalski, Jürgen Cox, and Matthias Mann. Expert system for computer-assisted annotation of ms/ms spectra. *Molecular & Cellular Proteomics*, 11(11):1500–1509, 2012.

[55] Wilhelm Lab. Spectrum fundamentals. https://github.com/wilhelm-lab/spectrum_fundamentals, 2023.

[56] Wassim Gabriel, Omar Shouman, and Mathias Wilhelm. PROSPECT PTMs Dataset - Multi-PTM. DOI:https://doi.org/10.5281/zenodo.11472525, 2024.

[57] Wassim Gabriel, Omar Shouman, and Mathias Wilhelm. PROSPECT PTMs Dataset - TMT. DOI:https://doi.org/10.5281/zenodo.8221499, 2023.

[58] Wassim Gabriel, Omar Shouman, and Mathias Wilhelm. PROSPECT PTMs Dataset - TMT-PTM. DOI:https://doi.org/10.5281/zenodo.11474099, 2024.

[59] Wassim Gabriel, Omar Shouman, and Mathias Wilhelm. PROSPECT PTMs Dataset - Test-PTM. DOI:https://doi.org/10.5281/zenodo.11477731, 2024.

[60] Richard D LeDuc, Veit Schwämmle, Michael R Shortreed, Anthony J Cesnik, Stefan K Solntsev, Jared B Shaw, Maria J Martin, Juan A Vizcaino, Emanuele Alpi, Paul Danis, et al. Proforma: a standard proteoform notation. *Journal of proteome research*, 17(3):1321–1325, 2018.

[61] Andy T Kong, Felipe V Leprevost, Dmitry M Avtonomov, Dattatreya Mellacheruvu, and Alexey I Nesvizhskii. Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature methods*, 14(5):513–520, 2017.

[62] Omar Shouman, Wassim Gabriel, and Mathias Wilhelm. PROSPECT on GitHub. https://github.com/wilhelm-lab/PROSPECT, 2023.

[63] Wilhelmlab - Computational Mass Spectrometry. prospect-ptms-irt (revision 8c1e8ed), 2024.

[64] Wilhelmlab - Computational Mass Spectrometry. prospect-ptms-ms2 (revision 91b3693), 2024.

[65] Wilhelmlab - Computational Mass Spectrometry. prospect-ptms-charge (revision 05f266f), 2024.

[66] Omar Shouman, Wassim Gabriel, and Mathias Wilhelm. DLOmix on GitHub. https://github.com/wilhelm-lab/dlomix, 2024.

[67] Mario Picciani, Wassim Gabriel, Victor-George Giurcoiu, Omar Shouman, Firas Hamood, Ludwig Lautenbacher, Cecilia Bang Jensen, Julian Müller, Mostafa Kalhor, Armin Soleymaniniya, et al. Oktoberfest: Open-source spectral library generation and rescoring pipeline based on prosit. *Proteomics*, 24(8):2300112, 2024.

[68] Ludwig Lautenbacher, Kevin Yang, Tobias Kockmann, Christian Panse, Matthew Chambers, Elias Kahl, Fengchao Yu, Wassim Gabriel, Dulguun Bold, Tobias K Schmidt, et al. Koina: Democratizing machine learning for proteomics research. *bioRxiv*, pages 2024–06, 2024.

[69] Shenheng Guan, Michael F Moran, and Bin Ma. Prediction of lc-ms/ms properties of peptides from sequence by deep learning*[s]. *Molecular & Cellular Proteomics*, 18(10):2099–2107, 2019.

[70] Vladimir Gorshkov and Frank Kjeldsen. Exploiting charge state distribution to probe intramolecular interactions in gas-phase phosphopeptides and enhance proteomics analyses. *Analytical Chemistry*, 96(3):1167–1177, 2024.

[71] Kaiyuan Liu, Chenghua Tao, Yuzhen Ye, and Haixu Tang. Specencoder: deep metric learning for accurate peptide identification in proteomics. *Bioinformatics*, 40(Supplement_1):i257–i265, 2024.

[72] Kevin Eloff, Konstantinos Kalogeropoulos, Oliver Morell, Amandla Mabona, Jakob Berg Jespersen, Wesley Williams, Sam PB van Beljouw, Marcin Skwark, Andreas Hougaard Laustsen, Stan JJ Brouns, et al. De novo peptide sequencing with instanovo: Accurate, database-free peptide identification for large scale proteomics experiments. *bioRxiv*, pages 2023–08, 2023.

[73] Julia Rechenberger, Patroklos Samaras, Anna Jarzab, Juergen Behr, Martin Frejno, Ana Djukovic, Jaime Sanz, Eva M González-Barberá, Miguel Salavert, Jose Luis López-Hontangas, et al. Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. *Proteomes*, 7(1):2, 2019.

[74] Ágnes Révész, Helga Hevér, Arnold Steckel, Gitta Schlosser, Dániel Szabó, Károly Vékey, and László Drahos. Collision energies: Optimization strategies for bottom-up proteomics. *Mass spectrometry reviews*, 42(4):1261–1299, 2023.

[75] Tobias Schmidt, Patroklos Samaras, Viktoria Dorfer, Christian Panse, Tobias Kockmann, Leon Bichmann, Bart Van Puyvelde, Yasset Perez-Riverol, Eric W Deutsch, Bernhard Kuster, et al. Universal spectrum explorer: a standalone (web-) application for cross-resource spectrum comparison. *Journal of proteome research*, 20(6):3388–3394, 2021.

[76] Celina Tretter, Niklas de Andrade Krätzig, Matteo Pecoraro, Sebastian Lange, Philipp Seifert, Clara von Frankenberg, Johannes Untch, Gabriela Zuleger, Mathias Wilhelm, Daniel P Zolg, et al. Proteogenomic analysis reveals rna as a source for tumor-agnostic neoantigen identification. *Nature communications*, 14(1):4632, 2023.

[77] Wassim Gabriel and Mathias Wilhelm. PrositPTM. *bioRxiv*, 2024.

[78] Chunwei Ma, Zhiyong Zhu, Jun Ye, Jiarui Yang, Jianguo Pei, Shaohang Xu, Ruo Zhou, Chang Yu, Fan Mo, Bo Wen, et al. Deeprt: deep learning for peptide retention time prediction in proteomics. *arXiv preprint arXiv:1705.05368*, 2017.

[79] Patrick Willems, Igor Fijalkowski, and Petra Van Damme. Lost and found: re-searching and re-scoring proteomics data aids genome annotation and improves proteome coverage. *Msystems*, 5(5):10–1128, 2020.

[80] Louise M Buur, Arthur Declercq, Marina Strobl, Robbin Bouwmeester, Sven Degroeve, Lennart Martens, Viktoria Dorfer, and Ralf Gabriels. Ms2rescore 3.0 is a modular, flexible, and user-friendly platform to boost peptide identifications, as showcased with ms amanda 3.0. *Journal of proteome research*, 2024.

[81] Di Zhang, Zhanyun Tang, He Huang, Guolin Zhou, Chang Cui, Yejing Weng, Wenchao Liu, Sunjoo Kim, Sangkyu Lee, Mathew Perez-Neut, et al. Metabolic regulation of gene expression by histone lactylation. *Nature*, 574(7779):575–580, 2019.

[82] Di Zhang, Jinjun Gao, Zhijun Zhu, Qianying Mao, Zhiqiang Xu, Pankaj K Singh, Cornelius C Rimayi, Carlos Moreno-Yruela, Shuling Xu, Gongyu Li, et al. Lysine l-lactylation is the dominant lactylation isomer induced by glycolysis. *Nature Chemical Biology*, pages 1–9, 2024.

[83] Charlotte Adams, Wassim Gabriel, Kris Laukens, Mario Picciani, Mathias Wilhelm, Wout Bittremieux, and Kurt Boonen. Fragment ion intensity prediction improves the identification rate of non-tryptic peptides in timstof. *Nature communications*, 15(1):3956, 2024.

[84] Tom Altenburg, Thilo Muth, and Bernhard Y Renard. yhydra: Deep learning enables an ultra fast open search by jointly embedding ms/ms spectra and peptides of mass spectrometry-based proteomics. *bioRxiv*, 2021.

[85] Muhammad Usman Tariq and Fahad Saeed. Specollate: Deep cross-modal similarity network for mass spectrometry data based peptide deductions. *PloS one*, 16(10):e0259349, 2021.

[86] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, and Ming Li. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature methods*, 16(1):63–66, 2019.

[87] Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, and Ming Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017.

[88] Noel M O'Boyle. Towards a universal smiles representation-a standard method to generate canonical smiles based on the inchi. *Journal of cheminformatics*, 4:1–14, 2012.

[89] Eleni E Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia E Kavraki. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry*, 6(1):132, 2023.

[90] European Organization For Nuclear Research and OpenAIRE. Zenodo, 2013.

[91] Wilhelm Lab. Spectrum io. https://github.com/wilhelm-lab/spectrum_io, 2023.

[92] Daniel B. Martin, Jimmy K. Eng, Alexey I. Nesvizhskii, Andrew Gemmill, and Ruedi Aebersold. Investigation of neutral loss during collision-induced dissociation of peptide ions. *Analytical chemistrys*, 77(15):4870–82, 2005.

[93] Stefka Tyanova, Tikira Temu, and Juergen Cox. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols*, 11(12):2301–2319, 2016.

[94] Daniel Paul Zolg, Mathias Wilhelm, Peng Yu, Tobias Knaute, Johannes Zerweck, Holger Wenschuh, Ulf Reimer, Karsten Schnatbaum, and Bernhard Kuster. PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration. *PROTEOMICS*, 17(21), 2017.

[95] Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nature communications*, 7(1):13404, 2016.

[96] Arnold Steckel, Katalin Uray, Gergo Kallo, Eva Csosz, and Gitta Schlosser. Investigation of neutral losses and the citrulline effect for modified h4 n-terminal pentapeptides. *Journal of the American Society for Mass Spectrometry*, 31(3):565–573, 2020.

[97] Martin R Larsen, Morten B Trelle, Tine E Thingholm, and Ole N Jensen. Analysis of posttranslational modifications of proteins by tandem mass spectrometry: Mass spectrometry for proteomics analysis. *Biotechniques*, 40(6):790–798, 2006.

[98] Michael D Hoffman, Matthew J Sniatynski, Jason C Rogalski, JC Yves Le Blanc, and Juergen Kast. Multiple neutral loss monitoring (mnm): A multiplexed method for post-translational modification screening. *Journal of the American Society for Mass Spectrometry*, 17:307–317, 2006.

[99] Xiaowei Zhao, Jiagen Li, Rui Wang, Fei He, Lin Yue, and Minghao Yin. General and species-specific lysine acetylation site prediction using a bi-modal deep architecture. *IEEE Access*, 6:63560–63569, 2018.

[100] Sian Soo Tng, Nguyen Quoc Khanh Le, Hui-Yuan Yeh, and Matthew Chin Heng Chua. Improved prediction model of protein lysine crotonylation sites using bidirectional recurrent neural networks. *Journal of proteome research*, 21(1):265–273, 2021.

[101] Xie-Xuan Zhou, Wen-Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si-Min He, and Zhifei Zhang. pdeep: predicting ms/ms spectra of peptides with deep learning. *Analytical chemistry*, 89(23):12690–12697, 2017.

[102] Jian Song, Fangfei Zhang, and Changbin Yu. Alpha-frag: a deep neural network for fragment presence prediction improves peptide identification by data independent acquisition mass spectrometry. *bioRxiv*, 2021.

[103] Hao Yang, Hao Chi, Wen-Feng Zeng, Wen-Jing Zhou, and Si-Min He. pnovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, 35(14):i183–i190, 2019.

[104] Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, and William S Noble. De novo mass spectrometry peptide sequencing with a transformer model. In *International Conference on Machine Learning*, pages 25514–25522. PMLR, 2022.

[105] Chunyuan Qin, Xiyang Luo, Chuan Deng, Kunxian Shu, Weimin Zhu, Johannes Griss, Henning Hermjakob, Mingze Bai, and Yasset Perez-Riverol. Deep learning embedder method and tool for mass spectra similarity search. *Journal of Proteomics*, 232:104070, 2021.

[106] Muhammad Usman Tariq and Fahad Saeed. Specollate: Deep cross-modal similarity network for mass spectrometry data based peptide deductions. *PloS one*, 16(10):e0259349, 2021.

[107] Wout Bittremieux, Damon H May, Jeffrey Bilmes, and William Stafford Noble. A learned embedding for efficient joint analysis of millions of mass spectra. *Nature Methods*, pages 1–4, 2022.

[108] Mostafa Kalhor, Joel Lapin, Mario Picciani, and Mathias Wilhelm. Rescoring peptide spectrum matches: Boosting proteomics performance by integrating peptide property predictors into peptide identification. *Molecular & Cellular Proteomics*, page 100798, 2024.