PROSPECT: Labeled Tandem Mass Spectrometry Dataset for Machine Learning in Proteomics

Omar Shouman * **Wassim Gabriel*** **Victor Giurcoiu** **Vitor Sternlicht**

Mathias Wilhelm

Computational Mass Spectrometry
School of Life Sciences
Technical University of Munich (TUM)
Freising, Germany
{firstname.lastname}@tum.de

Appendix

A Dataset Access

The dataset is hosted on Zenodo with a reserved DOI: https://doi.org/10.5281/zenodo.6602020 [1].

A dedicated GitHub repository provides utilities for the dataset https://github.com/wilhelm-lab/PROSPECT[2].

B Proteomics Terminology and Acronyms

- Retention Time (RT): The time a peptide spends in a column. It can also be defined as the time spent in the stationary and mobile phases
- 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 [3].
- MS/MS: Tandem Mass Spectrometry.
- Andromeda Score: A probabilistic score assigned to a spectrum by MaxQuant [4] to indicate the certainty of the identification. A higher score indicates higher confidence.

C Annotation Pipeline

As discussed in the paper, we used an expert system for annotation of the MS/MS spectra, which relies on domain-specific conditional rules. Table 1 lists the rules with their original name and number as described in the original expert system publication [5].

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.

*Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Table 1: Rules from the used expert system [5] for annotation of MS/MS spectra.

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

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 [6].

D Pre-processing for Machine Learning

Filtering

To filter our dataset, we remove all spectra with an andromeda score below a certain threshold. We used a score of 70 as a threshold to remove any uncertain identifications, resulting in less decoys in the dataset.

Data Split

As mentioned in the paper, we split the data based on the uniqueness of peptides sequences so that identical sequences do not end up in different splits. This is important to ensure that the same sequence, for example, is not in the two different splits (train and validation), resulting in less accurate evaluation of the trained models.

Indexed Retention Time

Current state of the art models only predict one value for retention time and not the complete elution profile. To prepare the data for training a model with a single retention time value, we get the best scoring peptide per raw file, then we get the median iRT value for peptides identified in multiple raw files. For predicting a single iRT, this approach gets the best possible representative iRT value for each peptide. We kept all the iRT values in the dataset so that it can also allow for predicting a complete elution profile for peptides.

Fragment Ions' Intensity

The final step in the fragment ions' pre-processing pipeline is to normalize the intensity values per spectrum. We scale all intensities in each spectrum by dividing them by the most intense annotated peak (max value) in the spectrum [7].

E Dataset File Format and Structure

We explained the overall schema of the data in Section 3. We provide further details on the file structure and description of the columns in the dataset.

The meta-data parquet file contains the following columns:

- `raw_file` Raw file name.
- `scan_number` Scan number of the MSMS spectrum.
- `modified_sequence` Sequence representation with the post-translational modifications.
- `precursor_charge` The charge of the precursor ion.
- `precursor_intensity` The intensity of the precursor ion.
- `mz [da]` The theoretical mass over charge of the peptide sequence.
- `precursor_mz [da]` Mass over charge of the precursor ion.
- `fragmentation` Type of fragmentation method used (HCD, CID).
- `mass_analyzer` Type of mass analyzer used to record the spectra (ITMS: Ion Trap, FTMS: Orbi Trap) .
- `retention_time [min.]` Uncalibrated RT in minutes where the MS/MS spectrum has been acquired.
- `indexed_retention_time` iRT calculated based on the Retention Time.
- `andromeda_score` Andromeda score for the associated spectrum.
- `peptide_length` Length of the peptide sequence without modifications.
- `orig_collision_energy` Collision energy used to acquire the spectra.
- `aligned_collision_energy` Calibrated collision energy.

Each annotation parquet file contains the following columns:

- `ion_type` Type of the fragment ion (y, b).
- `no` Number of fragment ion (1,n-1), where n is the peptide length.
- `charge` Charge of the fragment ion .
- `experimental_mass [da]` Mass of the experimental peak annotated.
- `theoretical_mass [da]` Theoretical mass of the fragment ion.
- `intensity` Normalized intensity of the peak.
- `neutral_loss` Molecules lost from fragment ion (Empty string if no neutral loss).
- `fragment_score` Score of the fragment ion to solve collisions.
- `peptide_sequence` Sequence representation with the post-translational modifications.
- `scan_number` Scan number of the MS/MS spectrum.
- `raw_file` Raw file name.

Note that the unit `Da` is the dalton or unified atomic mass unit.

F Further Exploratory Data Analysis

Table 2 provides summary statistics on the modifications that exist in the dataset. The reported counts represent the number of unique sequences in total and those with at least one modification. Moreover, counts for occurrences of the two modifications mentioned in Section 3 in the unique set of sequences are reported, where Carboxyamidomethylation and Oxidation correspond to UNIMOD IDs 4 and 35, respectively.

Table 2: Summary statistics of the modifications in the dataset.

Modified Unique sequences	Total Unique Sequences	Percentage	Carboxyamido-methylation	Oxidation
257 K	838 K	30%	167 K	165 K

G Differences to ProteomeTools and Experimentally-Acquired Spectral Libraries

PROSPECT provides value to proteomics and machine learning researchers by including several high-quality annotations and by being accessible in terms of format and structure for applying machine learning. The dataset provided is unfiltered without a score cutoff. Additionally, data from experimental sources, such as PRIDE, are not calibrated/aligned in terms of CE and RT. In order to reduce entry barriers, we provide a fully calibrated dataset, but also the option for researchers to come up with their own calibration methods since the raw uncalibrated values are provided as well. Additional data added to the resource will follow the same calibration and format; therefore, can be directly integrated into the existing dataset.

The dataset has the following additions/advantages over raw datasets and experimentally-acquired spectral libraries:

- annotations of MS/MS spectra via an expert system.
- annotations of neutral losses via an expert system.
- filtering of correct identifications based on confidence scores.
- aligned collision energy, calculated for each example of the dataset.
- indexed retention time, calculated for each example of the dataset.
- double annotations are available, researchers can explore different ways to resolve ambiguities on annotation.

Regarding accessibility, the dataset is prepared and structured to facilitate usage by machine learning researchers and practitioners without needing expert domain knowledge in proteomics. Some advantages include:

- easier access with parquet files, which can be easily read with common python libraries such as pandas.
- no need for special tools to read raw files generated from mass spectrometers.
- availability of metadata file with various features that can be additionally used while training models.
- traceability to raw data with a unique identifier.

H Elaboration on Supported Tasks

This section lists different tasks for which machine learning researchers can use our dataset directly. PROSPECT supports the tasks listed in Table 3 since it contains the required input-output mappings and additional features and annotations. We excluded retention time and intensity prediction since we elaborated on both tasks in the main manuscript.

I Experimental Details

We report more details on training Prosit [11] using our dataset for the two tasks; retention time prediction and intensity prediction. We used an Nvidia A30 GPU for training both models. Table 4 shows the duration of training, the loss function used, and the number of examples in each data split.

Table 3: Listing of tasks feasible with PROSPECT.

Name	Input	Target	Type	References
Retention Time	Sequence	RT	Regression (single value)	[8, 9, 10]
Retention Time	Sequence	RT	Regression (distribution)	-
Intensity prediction	Sequence	intensities	Regression (vector)	[11, 12, 13]
Fragment Presence	Sequence	Present/Not	Binary classification	[14]
Charge prediction	Sequence	Pre-dominant charge Charge distribution	Classification Regression	[15]
De novo sequencing	Intensity	Sequence	Classification/Ranking	[16, 17, 18, 19]
Sequence/Spectral Embedding	Sequence/Spectra	Embeddings	Representation/Similarity learning	[20, 21, 22]
Multiple properties	Sequence	RT, charge, intensity	Multi-task Learning	[15]
Sequence Clustering	Sequence	Cluster	Unsupervised Clustering	[23]

Table 4: Summary of experimental results from training Prosit.

Task	Epochs	Duration (hrs)	Loss	Metric	Train/Val/Test Split
Retention Time	325	2.5	MAE 2.63	Time Delta 85 seconds	500/200/200 k
Intensity	173	50	Spectral Distance 0.1272	Spectral Angle 0.8728	14/3.7/1.9 M

J Dataset Documentation: Datasheet for Datasets

J.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 a reference dataset, processed and curated for machine learning in Proteomics. Although the dataset is not constrained to specific tasks, the focus is on two common tasks in proteomics; retention time prediction and MS/MS spectrum 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.

J.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)?

Instances of the dataset are peptide sequences, their corresponding annotations, and meta-data for spectra. We have in Total 16.5B annotations for 5.7B unique peaks for 61M spectra of 716K unique peptides. Those are split in 12 packages with 983 pools. We uploaded 2 different file types for each package; 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 raw data [24], 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 [4] 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 two tasks we focused on, the targets are retention time (and indexed retention time) in the meta-data file and the annotated spectra in the annotation files.

Instaces of the dataset are generally linked together with the peptide sequences associated with each spectra. We use this sequence as a reference for annotation.

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 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 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.

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 [4]. 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 [25, 26, 27]. 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

J.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 [4] (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 have about 1000 peptides that we know 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 miss-identifications.

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 since the dataset we provide is based on ProteomeTools [24] and includes all examples.

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 [24] 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.

J.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.

Yes, as explained previously and in Sections 3 and D in the appendix.

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 [25, 26, 27].

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 miss-identifications.

J.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 [11, 28]

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 [1].

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, and assignment of annotations to peaks. For more details, please refer to section H.

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.

J.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 dataset to Zenodo with a DOI [1]. We also have a GitHub repository with utilities to download the dataset [2].

When will the dataset be distributed?

We published the dataset on Zenodo on the 9th of June 2022.

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.

J.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

Professorship for Computational Mass Spectrometry at the Technical University of Munich (TUM). The dataset is hosted on Zenodo [29]. Current maintainers are the authors and later other members of the Professorship at TUM. The ProteomeTools project web page will include a dedicated page for PROSPECT `https://www.proteometools.org/`.

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 miss-identifications or improve the quality of annotations, we will release subsequent versions with the respective updates. This will be versioned and announced in Zenodo 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 under the same DOI [1].

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.

K Author Statement

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

References

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

[2] Omar Shouman, Wassim Gabriel, and Mathias Wilhelm. Prospect. https://github.com/wilhelm-lab/PROSPECT, 2022.

[3] 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.

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

[5] 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.

[6] Wilhelm Lab. Spectrum fundamentals. https://github.com/wilhelm-lab/spectrum_fundamentals, 2022.

[7] Sven Degroeve, Niklaas Colaert, Joël Vandekerckhove, Kris Gevaert, and Lennart Martens. A reproducibility-based evaluation procedure for quantifying the differences between ms/ms peak intensity normalization methods. *PROTEOMICS*, 11(6):1172–1180, 2011.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Daniel P Zolg and Kuster Bernhard. ProteomeTools. `https://www.ebi.ac.uk/pride/archive/projects/PXD004732`, 2017. [Online; accessed 31-May-2022].

[26] Daniel P Zolg and Kuster Bernhard. ProteomeTools. `https://www.ebi.ac.uk/pride/archive/projects/PXD010595`, 2019. [Online; accessed 31-May-2022].

[27] Daniel P Zolg and Kuster Bernhard. ProteomeTools. `https://www.ebi.ac.uk/pride/archive/projects/PXD021013`, 2021.

[28] Mathias Wilhelm, Daniel P Zolg, Michael Graber, Siegfried Gessulat, Tobias Schmidt, Karsten Schnatbaum, Celina Schwencke-Westphal, Philipp Seifert, Niklas de Andrade Krätzig, Johannes Zerweck, et al. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nature communications*, 12(1):1–12, 2021.

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