



**HAL**  
open science

## Multienzyme deep learning models improve peptide de novo sequencing by mass spectrometry proteomics

Carlos Gueto-Tettay, Di Tang, Lotta Happonen, Moritz Heusel, Hamed Khakzad, Johan Malmström, Lars Malmström

► **To cite this version:**

Carlos Gueto-Tettay, Di Tang, Lotta Happonen, Moritz Heusel, Hamed Khakzad, et al.. Multienzyme deep learning models improve peptide de novo sequencing by mass spectrometry proteomics. PLoS Computational Biology, 2023, 19 (1), pp.e1010457. 10.1371/journal.pcbi.1010457 . hal-04403345

**HAL Id: hal-04403345**

**<https://hal.science/hal-04403345>**

Submitted on 18 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## RESEARCH ARTICLE

# Multienzyme deep learning models improve peptide *de novo* sequencing by mass spectrometry proteomics

Carlos Gueto-Tettay<sup>1</sup>, Di Tang<sup>1</sup>, Lotta Happonen<sup>1</sup>, Moritz Heusel<sup>1</sup>, Hamed Khakzad<sup>2</sup>, Johan Malmström<sup>1</sup>, Lars Malmström<sup>1\*</sup>

**1** Division of Infection Medicine, Department of Clinical Sciences Lund, Faculty of Medicine, Lund University, Lund, Sweden, **2** Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

\* [lars.malmstroem@med.lu.se](mailto:lars.malmstroem@med.lu.se)



## Abstract

Generating and analyzing overlapping peptides through multienzymatic digestion is an efficient procedure for *de novo* protein using from bottom-up mass spectrometry (MS). Despite improved instrumentation and software, *de novo* MS data analysis remains challenging. In recent years, deep learning models have represented a performance breakthrough. Incorporating that technology into *de novo* protein sequencing workflows require machine-learning models capable of handling highly diverse MS data. In this study, we analyzed the requirements for assembling such generalizable deep learning models by systematically varying the composition and size of the training set. We assessed the generated models' performances using two test sets composed of peptides originating from the multienzyme digestion of samples from various species. The peptide recall values on the test sets showed that the deep learning models generated from a collection of highly N- and C-termini diverse peptides generalized 76% more over the termini-restricted ones. Moreover, expanding the training set's size by adding peptides from the multienzymatic digestion with five proteases of several species samples led to a 2–3 fold generalizability gain. Furthermore, we tested the applicability of these multienzyme deep learning (MEM) models by fully *de novo* sequencing the heavy and light monomeric chains of five commercial antibodies (mAbs). MEMs extracted over 10000 matching and overlapped peptides across six different proteases mAb samples, achieving a 100% sequence coverage for 8 of the ten polypeptide chains. We foretell that the MEMs' proven improvements to *de novo* analysis will positively impact several applications, such as analyzing samples of high complexity, unknown nature, or the peptidomics field.

## OPEN ACCESS

**Citation:** Gueto-Tettay C, Tang D, Happonen L, Heusel M, Khakzad H, Malmström J, et al. (2023) Multienzyme deep learning models improve peptide *de novo* sequencing by mass spectrometry proteomics. *PLoS Comput Biol* 19(1): e1010457. <https://doi.org/10.1371/journal.pcbi.1010457>

**Editor:** Elena Papaleo, Danish Cancer Society Research Center, DENMARK

**Received:** August 3, 2022

**Accepted:** January 4, 2023

**Published:** January 20, 2023

**Copyright:** © 2023 Gueto-Tettay et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD037803. (<https://www.ebi.ac.uk/pride/archive/projects/PXD037803>) On the other hand, our custom version of DeepNovoV2 and all Snakemake workflows employed in this manuscript are stored in the Zenodo repository <https://doi.org/10.5281/zenodo.7155719>.

## Author summary

In recent years, the application of deep learning represented a breakthrough in the mass spectrometry (MS) field by improving the assignment of the correct sequence of amino acids from observable MS spectra without prior knowledge, also known as *de novo*

**Funding:** This work was supported by Foundation of Knut and Alice Wallenberg (2016.0023) to LH, JM and LM, Vetenskapsrådet 2020-02419 to JM and LM, and Alfred Österlunds Stiftelse to JM and LM. The funders had no role in study design, data collection, analysis, publication decision, or manuscript preparation. None of the researchers received a salary from the funders.

**Competing interests:** None of the listed authors have a conflict of interest.

MS-based peptide sequencing. However, like other modern neural networks, models do not generalize well enough as they perform poorly on highly varied N- and C-termini peptide test sets. To mitigate this generalizability problem, we conducted a systematic investigation to uncover the requirements for building generalized models and boosting the performance on the MS-based *de novo* peptide sequencing task. Several experiments confirmed that the training set's peptide diversity directly impacts the resulting model's generalizability. Data showed that the best models were the multienzyme models (MEMs), i.e., models trained from a compendium of highly diverse peptides, such as the one generated from digesting a broad of species samples with a group of proteases. The applicability of these MEMs was later established by fully *de novo* sequencing 8 of the ten polypeptide chains of five commercial antibodies and extracting over 10000 proving peptides.

This is a *PLOS Computational Biology Methods* paper.

## Introduction

Bottom-up mass spectrometry-based proteomics (MS) is focused on the sensitive identification and quantification of peptides and, thereby, proteins in arbitrarily complex samples [1,2]. In the standard workflow, peptides are first produced through the proteolysis of proteins with the enzyme trypsin. In the following step, the generated peptides are separated by liquid chromatography and measured by mass spectrometry in tandem (LC-MS/MS). Finally, the peptide-spectrum matches (PSM), the assignment of the peptide sequences to individual MS spectra, are produced using comprehensive compendia of reference protein sequences database [3].

Some of MS's remarkable applications are in the infection medicine proteomics field, where it is employed to characterize the molecular mechanism behind invasive bacterial diseases [4–6], modeling host-pathogen interactions [7–13] and investigate systemic proteome changes [14–18]. The use of the trypsin protease is justified by its efficiency, stability, and specificity to cleavage only at the C-terminal of the basic residues, arginine, and lysine [19]. However, its applicability is limited by the amino acid composition of the target proteins and the pH of the digestion solution [20,21]. Proteases other than trypsin, such as Elastase, Glu-C, Asp-N, Pepsin, ProAlanase, are employed to achieve different cleavage patterns or work in various pH ranges [22–25]. Despite the increasing maturity of bottom-Up MS, peptide identification is restricted to the sequences included in a reference database. Consequently, it is unattainable to study proteins derived from organisms without sequence or which are extinct, environmental samples, and microbiomes. Other examples involve therapeutic monoclonal antibodies, i.e., immune system proteins composed of heavy (HC) and light (LC) chains containing conserved and variable regions. The latter region is typically not contained in the traditional sequence databases for either chain [24,26,27]. To overcome this limitation, *de novo* MS peptide sequencing is intended to extract partial or complete sequence information directly from collected MS spectra. In this strategy, the identities and positions of the amino acids are determined by the differences in mass of a series of consecutive fragments, for example, fragment ions of type  $b$  and  $y$ . To this end, programs have been created which implement algorithms based on graph theory, Hidden Markov models, linear and dynamic programming, such as PEAKS [28], NovoHMM [29], Lutefisk [30], Sherenga [31], pNOVO [32,33], and PepNovo [34], among others. As in other fields of proteomics [35], the application of deep

learning represented a performance breakthrough in *de novo* MS peptide sequencing, as in the case of DeepNovo [27]. Deep learning algorithms attempt to simulate the behavior of the human brain—albeit by using many connected layers of neurons, which allows it to learn multiple levels of representation of high-dimensional data [35–38]. This key aspect translates into revolutionary advances in many research fields, such as image processing [39], speech recognition [40], and natural language processing [37]. In the supervised learning flavor, a model learns to make predictions based on labeled training data. Here, features like the amount of data and their diversity directly impact the resulting model's generalizability, i.e., their ability to react to new data and make accurate predictions. Therefore, generalizability is central to the success of a model and its further implementation [36,38]. DeepNovo software outperformed other state-of-the-art methods at the level of amino acids and peptides. It combines convolutional and recurrent neural networks and local dynamic programming to learn the characteristics of tandem mass spectra, fragment ions, and sequence patterns of peptides. A later version (DeepNovoV2) added an order-invariant network architecture (T-Net) and a sinusoidal *m/z* positional embedding [41], which exceeds its predecessor by at least 13% at the peptide level [42].

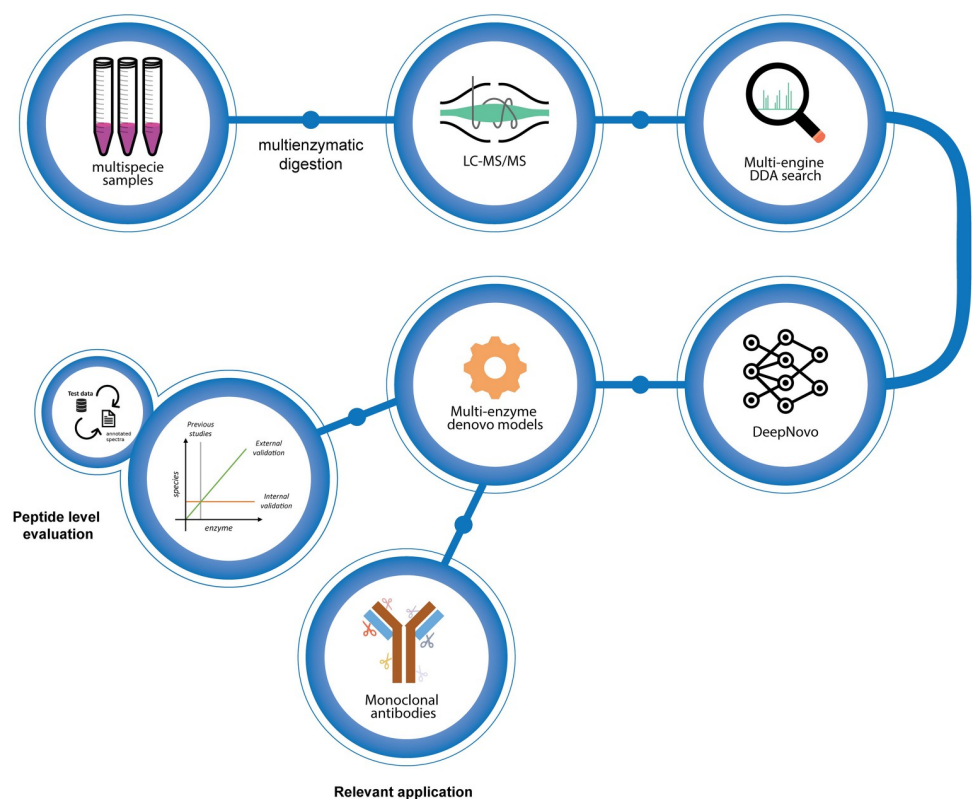
It has been reported that the generation and analysis of overlapping peptides through multi enzymatic digestion is an efficient procedure for tandem MS *de novo* protein sequencing [24,25,33,43]. Here, the same sample of the target protein is digested independently with a set of proteases with different cleavage patterns. Consequently, the generated peptides can overlap to reconstruct the primary structure of the protein of interest. This approach can even resolve some of the challenges encountered in conventional strategies, which depend on the cloning/sequencing of coding mRNAs [43–45]. Given the mentioned facts, integrating DeepNovo deep learning architecture to handle the multi enzymatic MS samples can be game-changing for the *de novo* protein sequencing field. In order to accomplish this, it requires generalized models. In this context, we refer to *de novo* sequencing models capable of successfully decoding the MS spectra of peptides with varied N- and C-terminus. Previous DeepNovo studies reported models trained exclusively from a compendium of tryptic peptides, referred to in this manuscript as trypsin-SEMs (trypsin Single Enzyme Models, [27,42]). This fact leaves the door open to questions related to the generalizability of the trypsin-SEMs. Firstly, it is uncertain whether these models have extended applicability to other MS datasets, i.e., having high accuracy on samples generated using proteases with different cleavages specificities to the one employed to produce the model's training set. In like matter, how the training set's composition impacts the resulting model's generalizability. Similarly, the effects of characteristics of the target spectra that facilitate peptide sequencing remain unexplored.

We studied the requirements for building generic DeepNovo models for the *de novo* MS sequencing task in the present work. For that purpose, we analyzed how the peptide composition and size of the training set affect the resulting model's generalizability. The efficiency of these *de novo* sequencing models was assessed on two highly sequence-diverse test sets by calculating the recall at the peptide level, i.e., the fraction of actual peptide sequences that were entirely correctly predicted [27,42]. Data showed reiteratively that using a collection of peptides with a wide variety of N- and C-termini amino acids led to 76% more generalizable models than the termini-restricted ones. Furthermore, DeepNovo models kept improving in the *de novo* peptide MS sequencing task as we continued extending the training set data with the multienzyme digestion of various species samples. We further proved the relevance of these multienzyme deep learning (MEM) models by *de novo* sequencing the heavy and light monomeric chains of five commercial monoclonal antibodies (mAbs). MEMs fully sequenced 8 of 10 target proteins, extracting over 10000 confirming and overlapping peptides from mAb MS samples digested with six different proteases. We consider that MEMs, combined with other

mass spectrometric techniques, will help *de novo* analyze MS samples of higher complexity, such as the mixture of mAbs.

## Results and discussions

To integrate DeepNovo into the *de novo* protein sequencing pipeline, we need deep learning models capable of performing *de novo* sequencing in MS spectra of samples digested with numerous proteases. Therefore, it is first mandatory to determine the basis for building such generic models. For that purpose, we explored the effect of the training set composition on the resulting model generalizability, following the workflow in Fig 1. We initially created five peptide datasets by digesting Detroit 562 cell line samples with five proteases: trypsin, chymotrypsin, elastase, gluc, and pepsin (see **Material and Methods** section for LC-MS/MS and spectra annotation details). In each dataset, 21492 peptides were randomly selected and split into training(90%), validation(5%), and test (5%) sets (see *De novo model generation and evaluation* section for details). We then systematically built multiple models from the training sets data. In order to assess all models' generalizability, it was essential to evaluate their performance on a dataset composed of highly variable peptides in terms of amino acid composition and peptide length distribution. For that reason, we constructed the Detroit test set by merging all five



**Fig 1.** We started with three sample cohorts; Detroit 562 cells, 5 commercially available antibodies, and a large collection of samples from different species. The samples were aliquoted and digested using five enzymes, measured using LC-MS/MS, and analyzed using traditional database searches with multiple search engines. All data were also analyzed using the published DeepNovo deep learning model. Several DeepNovo models were created, see text for details, and evaluated in three ways. The internal validation evaluated the model performance on data generated with the same enzyme(s) as the model was trained with. The external validation evaluated the model performance using data generated with enzyme(s) different from the model creation data. We finally assessed each model's performance in *de novo* sequencing five full-length antibodies.

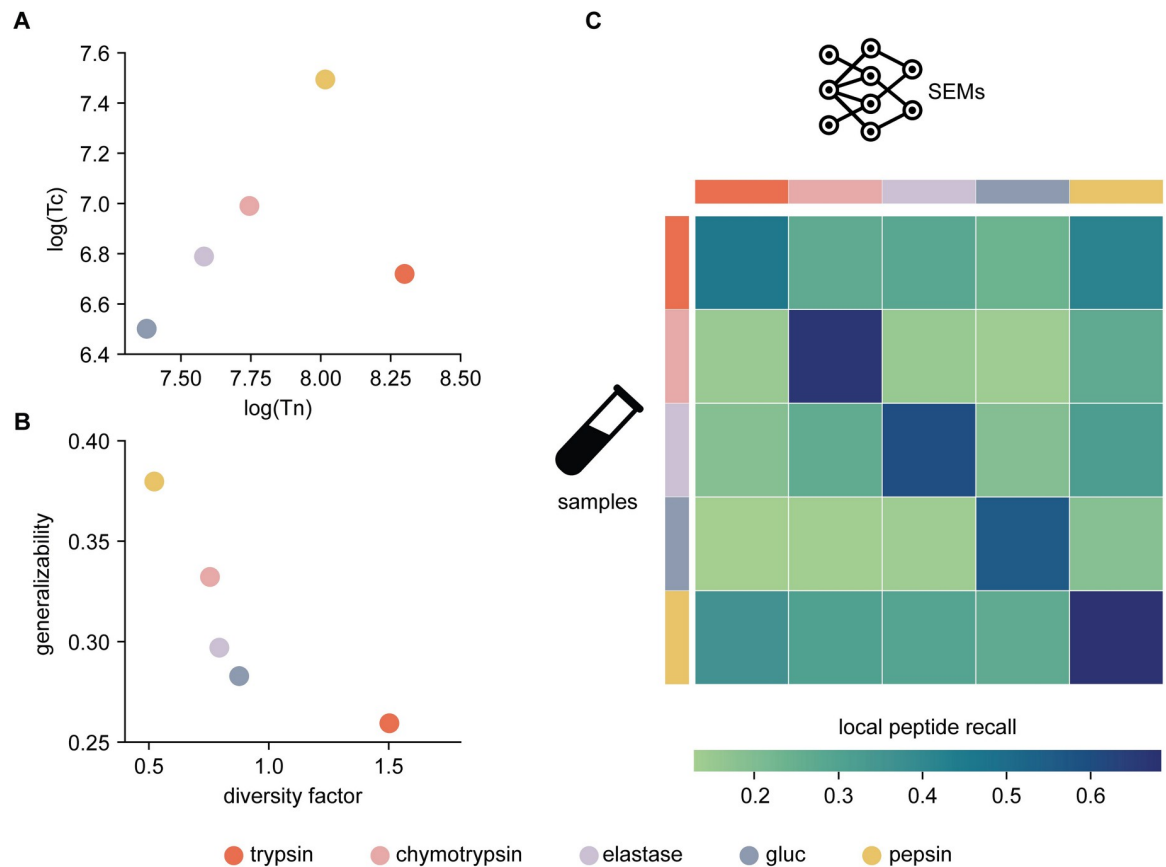
<https://doi.org/10.1371/journal.pcbi.1010457.g001>

test sets. Here, we used the recall at the peptide level as a quantitative and comparative metric of all trained models' capability for successfully *de novo* sequencing peptides with varied N- and C-terminus. Following that logic, we used the peptide recall on the complete test sets as a metric for the generalizability assessment (global peptide recall). Similarly, we calculated the peptide recall on the Detroit set's components, i.e., local peptide recall. In addition, given that the protease employed during the sample preparation has a direct effect on the resulting peptides collection termini variability, we calculated the number of unique trimers on the N-terminal (Tn) and C-terminal (Tc) for all the generated models' training sets in this study. Using trimers leads to unique and not overlapping termini N/C-fragments, as the minimum peptide length for the DDA search was set to 6. Likewise, Tn and Tc values increase with the model size. Overall, Tn and Tc are quantitative metrics for the extent of the training sets' variability at each peptide termini. Higher values of Tn and Tc represent higher variability in the peptide dataset at N and C-termini, respectively. Moreover, selecting trimers allowed us to measure the termini variability dependent on the model's size. We also introduced the diversity factor (DF), defined as  $\log(Tn/Tc)$ , as a measure of the variability balance between the training set's N- and C-terminus. DF values near zero represent models with a better balance between the number of trimers at each terminal. Similarly, positive and negative DF values indicate a larger proportion of Tn and Tc, respectively. The [S1 Table](#) includes all generated models' diversity attributes and performance on the Detroit test set.

### Nonspecific enzymes training datasets yield more generalized models

We built the first round of models from the five individual enzyme datasets and identified them as Single Enzyme Models (SEMs). [Fig 2](#) displays the characteristics and performance of all five SEMs on the Detroit test set. Two findings are worth mentioning regarding SEMs: 1) Using less specific proteases for the peptide generation leads to more N/C termini balanced training sets ([Fig 2A](#)). In contrast to pepsin, trypsin protease has a high specific cleavage pattern that generates a training set with high Tn and low Tc values, as the peptides end in either arginine or lysine amino acids. This observation is supported by the DF values for SEMs, i.e., pepsin (0.52) < chymotrypsin (0.76) < elastase (0.79) < Glu-C (0.88) < trypsin (1.50); 2) Models' generalizability correlates inversely with DF values ([Fig 2B](#)). When comparing performance on the Detroit test set at the peptide level, data shows that SEMs built with less specific enzymes, specifically pepsin-SEM, chymotrypsin-SEM, and elastase-SEM, outperform 14–46% of those generated from proteases with more specific cleavage patterns, such as gluc-SEM and trypsin-SEM. These differences in SEMs' generalizability are explained when considering their local peptide recall on the Detroit set's components ([Fig 2C](#)). We found that the most contributing factor was related to the models' performance on inter-enzyme datasets, e.g., where the proteases for generating the training and test sets differed. An illustrative example arises when considering the local peptide recall on the chymotrypsin, elastase, and gluc peptide datasets, for which the performance of the pepsin-SEM was approximately 46–86% higher than the trypsin-SEM. In addition, all SEMs performed best when there was a match between the protease employed to generate the SEM's training set and the Detroit set's portion. In these cases, local peptide recall ranged from 0.46 to 0.69. These results are comparable to previous Deep-Novo works where only trypsin was used [[27,42](#)]. Here, local peptide recall values show that less specific SEMs outperformed 6–48% of the highly cleavage pattern-specific ones. These results suggest that SEMs generated from the digestion with trypsin and gluc are more biased at the spectra decoding stage, especially for purposing the C-terminus peptide amino acids.

Inspired by the results of the first round, we then decided to test if it was possible to modulate models' generalizability as a function of their training set's diversity factor. For that

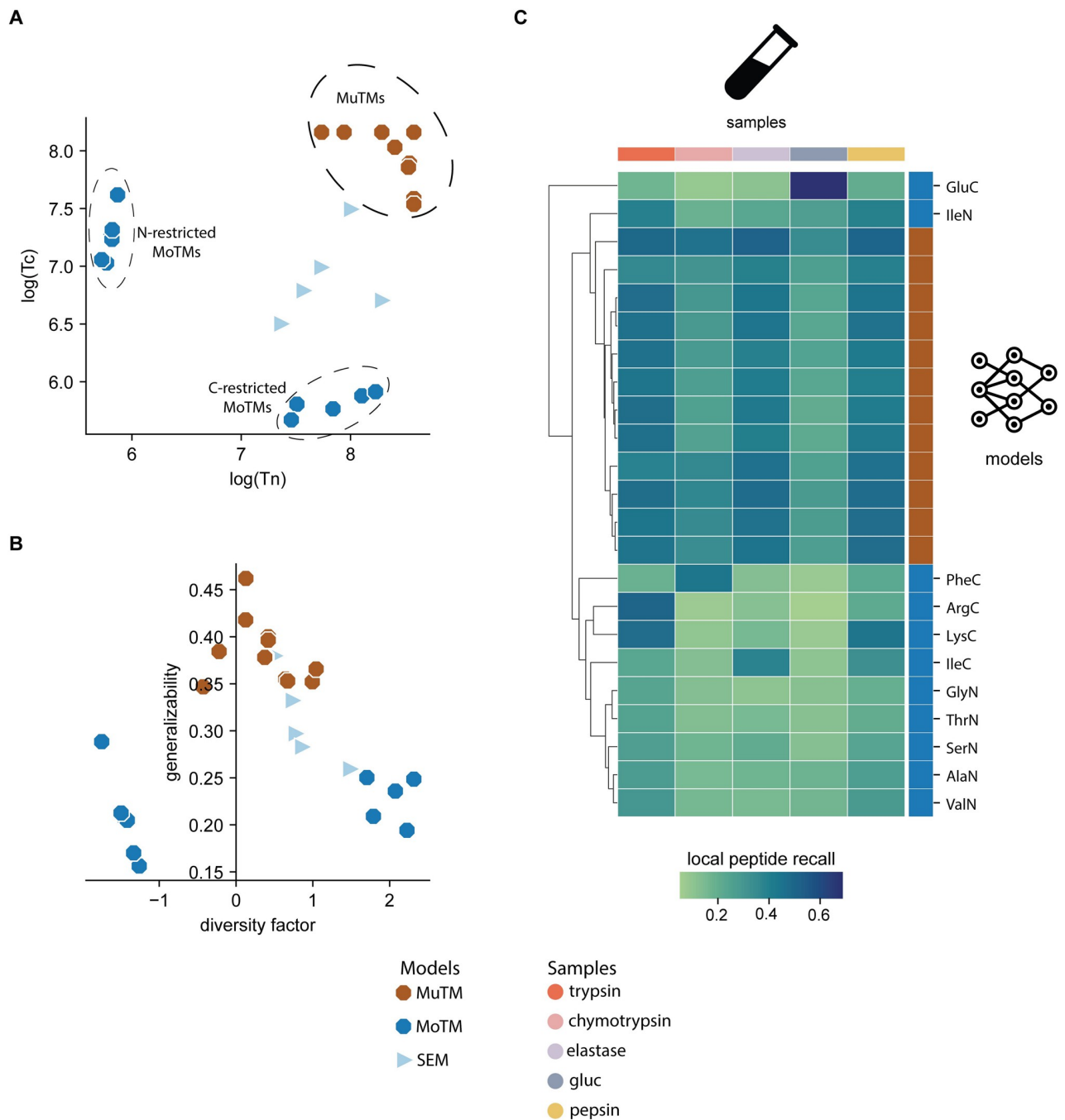


**Fig 2.** Single Enzyme models (SEMs) performance on Detroit test set: A) Tn and Tc values for SEM training sets; B) SEMs' generalizability vs. training set' diversity factor C) SEMs peptide performance on the individual enzyme-specific datasets composing the Detroit test set. The color scheme for both models and samples is at the bottom.

<https://doi.org/10.1371/journal.pcbi.1010457.g002>

purpose, we built new models distributed in two categories: 11 monoterminal (MoTMs) and 12 multiterminal (MuTMs) models. In MoTMs, the training sets were restricted by using peptides that share one specific amino acid at one of the termini positions. Given the MS data available, we built AlaN, GlyN, GluC, IleC, IleN, PheC, ArgC, LysC, ThrN, SerN, and ValN MoTMs. The models' nomenclature is composed of the amino acid three-letter code followed by the termini type; for example, in the ThrN and PheC MoTMs' training sets, all peptides have a Thr or Phe amino acid at the N- or C-terminal, respectively. Contrary to MoTMs regarding the diversity factor feature, MuTMs prioritized maximum variability at both terminals by selecting peptides from all SEMs' training sets. Furthermore, for a fair comparison with the previous SEMs' global and local peptide recall results, MoTMs and MuTMs were built with the same amount of spectra as SEMs. [Fig 3](#) displays MoTMs and MuTMs characteristics and performance on the Detroit test set.

Considering SEMs as reference, three new groups are distinguishable regarding Tn and Tc values distributions ([Fig 3A](#)). Two groups belong to MoTMs, which have low Tn and Tc values for the N-termini and C-termini restricted MoTMs, respectively. The third group belongs to the MuTMs, containing high values for both Tn and Tc parameters. [Fig 3B](#) shows that MuTMs are more termini-balanced and generalized than all MoTMs ones. According to the mean values of the global peptide recall on the Detroit test set, MuTMs outperform 76% of MoTM models. Moreover, half of the MuTMs generalize better than the pepsin-SEM, while



**Fig 3. Modulating models' generalizability by varying their training sets' termini variability.** Comparing Monoterminal (MoTM) and Multiterminal (MuTM) models' characteristics and performances on the Detroit test set: **A.** Training sets N/C-termini variability for MoTMs and MuTMs; **B.** Models' generalizability as a function of their termini diversity factor; **C.** MoTMs' and MuTMs' local peptide recall on Detroit test set components. We included SEMs' Tn, Tc, and diversity factor parameters as a quantitative reference. Color and shape schemes for models and sample types are at the bottom.

<https://doi.org/10.1371/journal.pcbi.1010457.g003>

the other half of the models were better than the chymotrypsin-SEM one. In contrast, 10 of 11 MoTMs-biased were worse than the trypsin-SEM at generalizing. On the other hand, the models' performance on the Detroit test set's components shows how MuTMs cluster together as they exhibited more uniform local peptide recall values across all sample types (Fig 3C). On



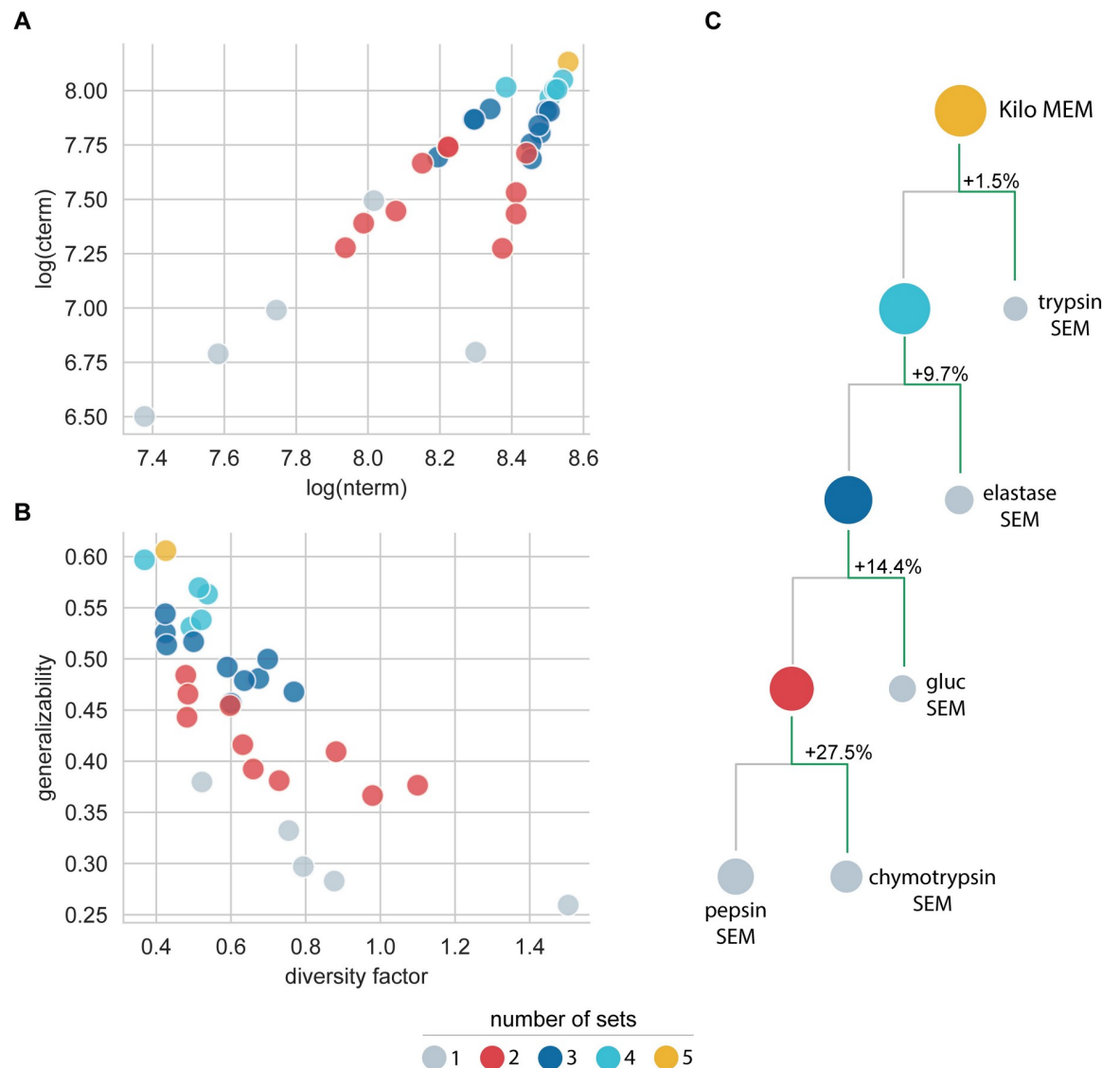
the contrary, the MoTMs' performance depended on the cleavage rules' overlap between the sample and the model's training set. For example, the ArgC-MoTM performed best on the trypsin sample. However, local peptide recall values dropped 57–90% in the remaining sample types. Similar behavior in performance was observed in other MoTMs, such as GluC-MoTM and PheC-MoTM. These observations suggest that, under the same amount of training data, it is possible to design more generalizable models by maximizing and balancing the training set's Tn and Tc values.

### Large MEMs perform best

Since all SEMs perform the best on similar data types as the model's training set, we then decided to build 26 new models by mixing all possible combinations of the five Single Enzyme models' training sets, i.e., multienzyme models (MEMs) from the combination of 2 (10 MEMs), 3 (10 MEMs), 4 (5 MEMs), and 5 (1 MEM) SEM-datasets. Here, the MEM composed for all five Detroit 562 peptide datasets was called the Kilo MEM. Data shows that appending one or more different peptide datasets to any existing SEM dataset yields growth in Tn, Tc, and generalizability parameters for the resulting MEM (Fig 4). As expected, the increase in Tn and Tc values was more noticeable when the merged datasets did not share the same cleavage rules as in chymotrypsin—gluc and trypsin—elastase—gluc dataset combinations (Fig 4A). Furthermore, generalizability and diversity factor values suggest that MEMs generalize better and are more termini-balanced as we increase the number of peptide datasets (Fig 4B). An illustrative example of MEMs' rising performance is shown in Fig 4C, where we displayed the path to generating the Kilo MEM from the pepsin-SEM. Two observations are worth mentioning: 1) new datasets contributed positively to the resulting MEMs' generalization, and 2) the formed MEM always performed better than its antecessors models. The Kilo MEM not only doubles the termini peptide dataset variability but also produced an increase of 38% in diversity factors concerning all SEMs. As a result, the Kilo MEM outperforms 1.8–2.4 times the SEMs.

The results of the SEM and MEMs demonstrated that features such as the training set's size and peptide sequence variability significantly impact the resulting model's generalizability. At this point, we hypothesized that expanding sequence variability by creating a training set that includes peptides across different species will lead to a more generic model than the Kilo MEM. To prove it, we generated an external dataset, called here Giga, by digesting various species samples, such as *Saccharomyces cerevisiae*, *Escherichia Coli*, *Equus caballus*, *Streptococcus pyogenes*, and *Mus musculus* with trypsin, chymotrypsin, elastase, and gluc proteases. We followed the same protocol for sample injection, MS detection, and database search (See [Material and methods](#)). After spectra annotation, the Giga dataset was ten times larger than the Detroit 562 dataset. We then trained and applied the Giga MEM to the Detroit test set and compared the results with the Kilo MEM. Data shows that the Giga MEM generalized 29.4% better than the Kilo MEM, outperforming 24–41% in all Detroit test set's composing sample types (Fig 5). In the same way, the Giga MEM generalizes 2.1–3.0 times better than the SEMs.

The Giga dataset was also used as an external test set. Specifically, we tested the generalizability of the 5 SEM and 26 MEMs. Interestingly, generalizability values on the Giga test set supported our previous findings on the best conditions to build more generic models ([S1 Text](#)). Here, it is crucial to mention the pepsin-SEM results; In the Detroit test set case, the most considerable portion of *de novo* sequenced spectra corresponded to peptides generated with the same protease as the SEM's training set. However, pepsin was not part of the multienzyme protocol for generating the Giga external peptide test set. Despite that, the pepsin-SEM performed best among all SEMs. Overall, generalizability results on the Detroit and Giga test



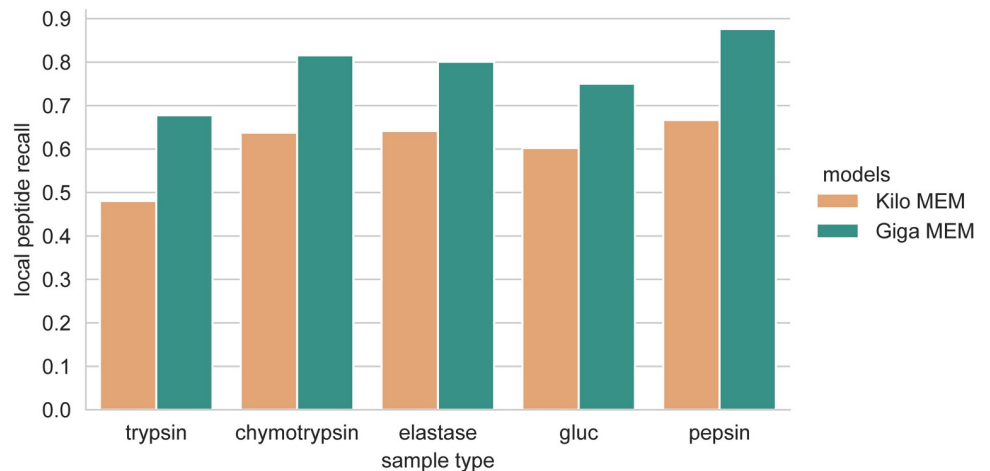
**Fig 4. Characteristics and performance of multienzyme (MEM) models.** **A.** Training sets' Tn and Tc values; **B.** MEMs' generalizability vs. diversity factor; **C.** Sequential building of Kilo MEM from all five SEM datasets. The size of circles is proportional to their generalizability values on the Detroit test set. The color scheme at the bottom reflects the models' characteristics variation with the number of combined datasets. SEM data (displayed in gray) were used as reference.

<https://doi.org/10.1371/journal.pcbi.1010457.g004>

sets suggest that, like other deep learning architectures, DeepNovo kept improving in the *de novo* peptide MS sequencing task as we fed the model with extensive and highly diverse peptide MS data.

### Fragment ions distribution impact MS *de novo* peptide sequencing

After establishing the criteria for building generalizable models, we further explored how the peptide composition impacts the ability to *de novo* sequence its spectrum correctly. In this respect, we studied the Kilo MEM results on the Giga test set (Fig 6). Initially, we evaluated the effect of the peptide length distribution on the overall deep learning model's performance by



**Fig 5. Comparing the performance of the Giga and Kilo MEMs across the Detroit test set's different sample type components.**

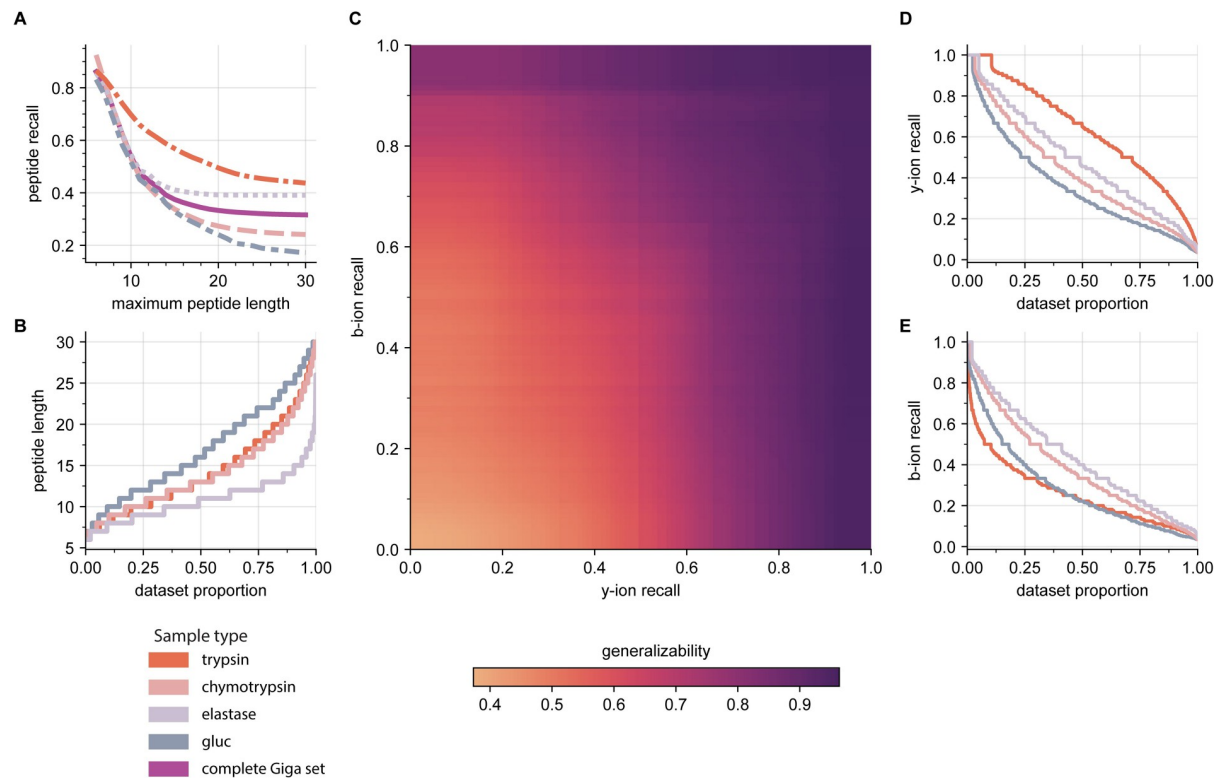
<https://doi.org/10.1371/journal.pcbi.1010457.g005>

tracking the peptide recall as we varied the maximum peptide length (Fig 6A). We observed that performance decreased as we included longer peptides in the test set. Data shows that the probability of *de novo* MS sequencing correctly 6-residue peptides was 86.1% and fell quickly to 40% when considering peptides of up to 14 residues.

Moreover, this performance decay differed for all components of the Giga set, suggesting that the identity of the peptides also impacts their chance of being MS sequenced. To explain these differences across the four datasets, we calculated the peptide length distribution (Fig 6B). Data shows that 75% of data in the elastase dataset are peptides of length 12 or shorter, explaining why it was more accessible to *de novo* MS sequence elastase data over chymotrypsin and gluc data. For the latter, 75% of the data were peptides of length 13 or longer.

Since the peptide length distributions could not explain performance differences related to the trypsin sample, we further calculated singly-charged b- and y-ion recall for all peptides spectra composing the Giga test set, e.g., the proportion of the fragment ions found experimentally over the total expected ones theoretically. Here the ion recall is a quantitative metric of the ability of a particular peptide to produce b/y-ions under specific experimental conditions [24,25,46,47]. For the fragment ions extraction, the m/z tolerance was 15ppm. We also calculated the peptide recall as a function of the minimum values for the b/y-ion recall pairs.

The b/y-ions recall grid shows that the probability of *de novo* MS sequence correctly a peptide increase with its capacity of producing either b- or y-ions (Fig 6C). Data shows that the global peptide recall on the Giga test set was higher than 70% when peptides produced at least 80% and 60% of the expected b- and/or y-ion fragments. These results suggest that the *de novo* MS sequencing performance on a specific sample type is bound to its b/y ion recall distributions. Fig 6D shows that the y-ion recall distribution order fits the peptide recall behavior for all sample types. It is worth mentioning that the tryptic peptides had the highest proportion of the expected singly-charged y-ions compared to the other sample types, explaining its remarkable performance across a wide range of peptide lengths (Fig 6A), i.e., 55% of the annotated spectra had at least 60% of the y-ions expected. For these peptides, y-ion fragments bear a charged residue, like arginine or lysine, which are more abundant and produce more intense peaks under the HCD fragmentation method [48,49]. On the contrary, the peptides from the digestion with gluc had a low proportion of y- and b-ions (Fig 6E). Furthermore, elastase b/y-ion recall distributions are consistent with a high proportion of short peptides.

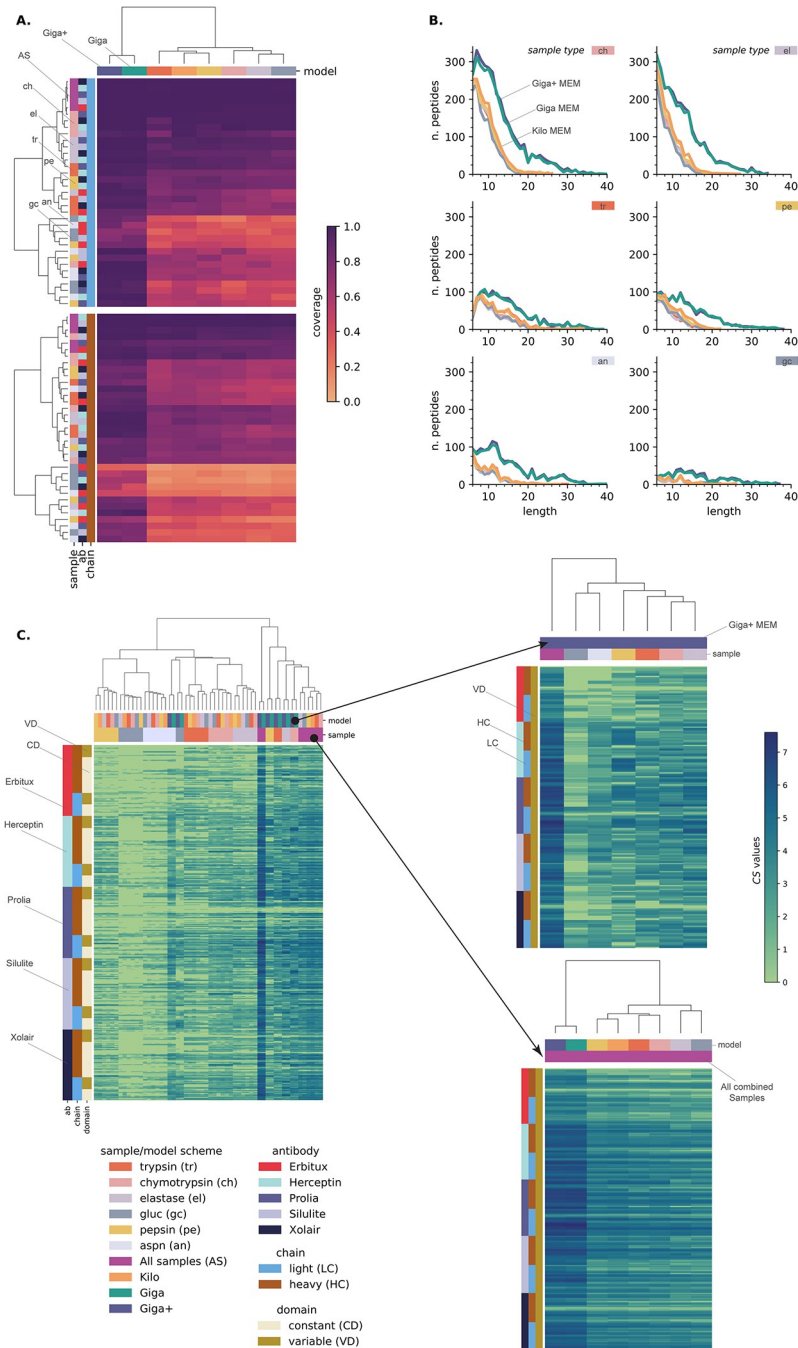


**Fig 6. Kilo MEM *de novo* sequencing results on the Giga external test set.** A. peptide recall as a function of the maximum peptide length in the test set at a global and local level; B. peptide length distribution for all the Giga test set's sample types; C. generalizability as a function of the minimum singly-charged b-y ion recall grid values; D. y-ion recall and E. b-ion recall distributions for the trypsin, chymotrypsin, elastase, and gluc sample types spectra. The color scheme for the sample types is at the bottom.

<https://doi.org/10.1371/journal.pcbi.1010457.g006>

## MEMs for full-length *de novo* sequencing of antibodies

Once we established the requirements for building generalizable models and how the quality of the input spectra impacts the subsequent *de novo* MS peptide sequencing process, we tested the efficiency of using the MEMs in the *de novo* protein sequencing pipeline. For this effort, we selected a challenging and biological interest system, such as the complete sequencing of monoclonal antibodies (mAbs). We aimed to fully *de novo* MS sequence the heavy (HC) and light (LH) chains of five commercial mAbs: Erbitux, Herceptin, Prolia, Silulite, and Xolair. We digested each mAb sample with six proteases: trypsin, chymotrypsin, elastase, gluc, pepsin, and aspn. It is worth mentioning that the latter enzyme was not part of the models' generation protocol. On the other hand, we created the Giga+ MEM by combining the training sets of the Kilo and Giga MEMs. We considered eight models (5 SEM + 3 MEMs) for comparison purposes. For analyzing results, we initially calculated the relative coverage for the entire variables space, i.e., models x samples x chains matrix (Fig 7A, S2 Text, S1 Table). This way, we got an insight into the model performance across all sample types and which enzymes facilitate the *de novo* sequencing of the HC and LC subunits. In addition, we examined the length distribution of the sequence matching peptides for all sample types (Fig 7B). These plots provide information about the decoding power of the models. It also shows the capacity of the different proteases to produce easily detectable peptides from the *de novo* MS sequencing perspective. Here, we initially discussed the impact of using different proteases for the *de novo* sequencing of monoclonal antibodies, specifically, how it affects the ability of all models to achieve high protein coverage and generate a high number of different easy-to-decode peptides. We then



**Fig 7. Complete *de novo* sequencing of commercial monoclonal antibodies by deep learning models.** A) coverage of all the light (top) and heavy (bottom) chains for all types of samples. B) length distribution for all matching peptides extracted from each type of sample; C) Confident positional score (CS) values for all proteins' variable and constant domains. General view of CS values for the evaluated models on all sample types (Left panel). Zoom into the CS values for the heavy and light chains variable domains for the Giga+ deep learning model across all sample types (Right top panel) and all models when combining all sample types (Right bottom panel).

<https://doi.org/10.1371/journal.pcbi.1010457.g007>

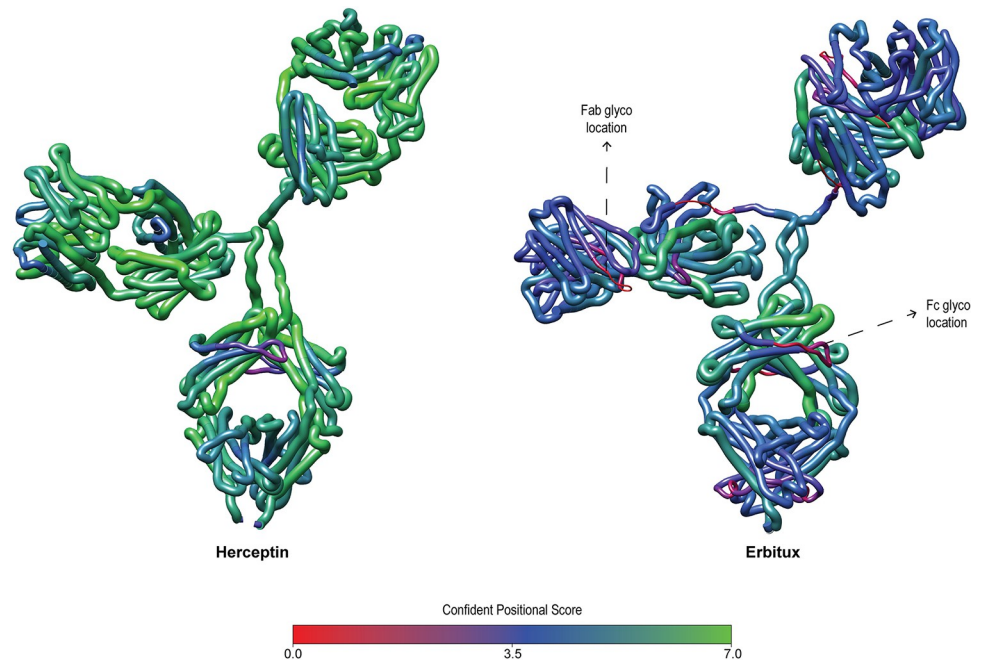
examined the individual performance of MEMs against SEMs across all sample types. Finally, we showed the capacity of MEMs for producing a high amount of overlapped peptides across the variable and constant domains of the studied proteins.

Regarding the sample types, the data shows that working with the chymotrypsin and elastase proteases had many benefits related to good protein coverage (Fig 7A) and the extraction of a high amount of matching peptides (Fig 7B). Data shows that digesting the samples with these proteases yields better individual protein coverage, wherein in 75% of cases, the sequence relative coverage was at least 0.80 and 0.75 for chymotrypsin and elastase, respectively. Additionally, the total amount of peptides extracted was 2 to 8 times greater than the rest of the proteases (S2 Text). It is worth noting that these were the only enzymes where, for lengths between 6 and 9, all of the considered deep learning models identified more than 100 peptides. These observations suggest that working with the chymotrypsin and elastase proteases leads to high amounts of readily *de novo* MS sequenceable peptides. As expected, the gluc and aspn digested samples got the lowest matching peptide extraction values, yielding the worst individual protein coverages. These proteases produced long peptides with low b- and y-ion recalls, making them more difficult to *de novo* sequence.

When comparing the performance of the deep learning models, the Giga and Giga+ MEMs were evident superior after considering the values of the protein coverages and the amount of matching extracted peptides parameters. For the Giga+ MEM, the median value of protein coverage was 0.96 after considering all mAbs and sample types. Moreover, it extracted 10367 unique and confirming peptides, an amount 2–2.8 times greater than the Kilo MEM and all SEMs (S2 Text). Interestingly, and based on the same parameters, the pepsin SEM was among the five SEMs. These findings supported our previous statements about the necessary criteria for building generalizable models. It is worth noting that the Giga+ MEM sequenced all light chains and 3 of 5 mAbs heavy subunits for the combined sample results, i.e., Herceptin, Silulite, and Xolair mAbs. The remaining proteins had coverage of at least 0.97. It is essential to consider that, in mAb, the HC subunit can bear glycans in their constant region [50,51]. In some cases, such as for Erbitux, glycans are also found in the HC variable region [52].

As the overlapping of peptides is necessary for the assembly of protein sequences, we also decided to go deeper into the analysis of MAb *de novo* results and introduce the *confident positional score* (CS). For a residue in the position  $i$  of the protein sequence, is defined as  $C_i = \log_2(f_i+1)$ . Here  $f_i$  is the positional frequency for position  $i$ , i.e., the number of *de novo* sequenced matching peptides for position  $i$  in the protein sequence (Fig 7C). Higher consecutive CS values represent regions with more evidence in the *de novo* protein sequencing process, being especially important for MAb HC and LC variable regions, for which the sequences are unknown. In contrast, sequence regions with no detected peptides have a zero positional frequency, ergo, a zero CS value. After combing all sample types, the Giga+ MEM got a positional frequency greater than ten for 90.7% of the amino acids comprising the study mAbs. Moreover, this parameter value increased to 50 or more for 45.7% of said amino acids. Similarly, there were no confirming peptides for only 0.03% of residues. Furthermore, For the mAbs variable region, the median positional frequency was 45 and 51 for the HC and LC subunits, respectively (S3 Text). For the five HC subunits, data show that CS values decreased up to 30% in the glycans' surrounding regions, likely because of a steric effect as these bulky species prevent efficient digestion. In the case of the Erbitux mAb, the regions with zero CS values matched the glycans location for the HC constant and variable domains (Fig 8), suggesting that removing the glycans should be incorporated in the sample preparation to guarantee the complete MS sequencing of mAbs. Given the coverage and positional frequency results, the findings discussed here set a precedent for using multienzymatic deep learning models as an alternative for sequencing proteins from their multienzymatic digestion.

In future studies, it may be interesting to explore using the multienzyme *de novo* sequencing protocol in conjunction with other complementary MS techniques like Top-down to



**Fig 8. 3D segmented worm representation of the mAbs with the highest (Herceptin) and lowest (Erbitux) *de novo* MS sequencing CS values.** The thickness and color of the protein chains are proportional to their CS values.

<https://doi.org/10.1371/journal.pcbi.1010457.g008>

sequence mixtures of mAbs. Similarly, the positional frequency concept opens room for developing new MS-based protein assembly methods.

## Conclusions

We proposed the use of MEMs to improve the *de novo* sequencing of peptides and proteins from DDA-MS data. Toward that aim, the effects of the properties of the training and test sets on the *de novo* sequencing process were explored. On the one hand, the data suggest that variability at both terminals, among the peptides which make up the training set, affects eventual generalizability. Consequently, the use of multiple proteases is recommended to generate more robust models. In the same vein, since DeepNovoV2 learns characteristics of spectra and sequences, an increase in the number of data points also improves the resultant model performance. These claims are supported by the peptide recall results for the test sets and the number of peptides extracted from the samples produced by multienzymatic digestion of commercial antibodies. On the other hand, it was discovered that the models' *de novo* sequencing capacity is limited by the identity of the peptides and experimental conditions, which have direct consequences on their ability to produce ionic fragments of interest. This result explains why peptide recall fell with an increase in length of peptides, as well as the differences found among the samples from trypsin, chymotrypsin, elastase, and gluc. Finally, the findings described here will assist in other areas of peptidomics, the creation of Data-Independent-Acquisition libraries, and the sequencing of complex mixtures of monoclonal antibodies.

## Material and methods

### Sample preparation for mass spectrometry

**Commercial antibodies.** For sample preparation for mass spectrometry of commercial antibodies, 10  $\mu$ g of each (Xolair, Novartis; Herceptin, Roche; SiLuLite, Sigma MSQC4

Universal Antibody Standard; Prolia, Amgen; and Erbitux, Merck) was denatured with 8M urea—100 mM ammonium bicarbonate, the disulphide bonds reduced with 5 mM Tris (2-carboxyethyl) phosphine hydrochloride (TCEP) for 60 min at 37°C, 800 rpm, and alkylated with 10 mM iodoacetamide for 30 min in the dark at room temperature. The samples were diluted to a urea concentration <1.5 M with 100 mM ammonium bicarbonate. The antibodies were digested separately with 1 µg of trypsin, Promega; chymotrypsin, Promega; LysC/trypsin, Promega; elastase, Promega; GluC, Promega; or AspN, Promega for 18h at 37°C, 800 rpm. The digested samples were acidified with 10% formic acid to a pH of 3.0. The peptides were purified and desalted using SOLAµ reverse phase extraction plates (Thermo Scientific) according to the manufacturer's instructions. Peptides were dried in a speedvac and reconstituted in 2% acetonitrile, 0.2% formic acid prior to mass spectrometric analyses.

**Detroit 562 cell line.** Briefly, ca.5 million cultured mammalian epithelial cells (Detroit 562 cell line) were kindly provided by Sounak Chowdhury. Suspension cells were first centrifuged at 5000g rcf, 4°C for 10 mins, followed by aspiration of supernatant and one time cold 1X PBS wash. Remained cell pellets were then added with 1 ml lysis working solution, composed of 1X RIPA lysis and extraction buffer, ThermoFisher, and 1X protease/phosphatase inhibitor cocktail, ThermoFisher. After 15 min incubation on ice, the cell lysates were precipitated by trichloroacetic acid (TCA), washed with 3X acetone, and dried in a speedvac. Completely dried protein extracts were reconstituted in 100 mM ammonium bicarbonate buffer and measured for protein concentration by BCA assays, ThermoFisher. 10 µg of cell lysate proteins were aliquoted for each reaction, 10 experimental replicates for each enzyme, 5 enzymes in total. Sample preparation of reduction, alkylation, enzyme digestion, acidification was described as above. Specifically, C18 spin columns were used for purification and desalting after the digestion. Except for the pepsin-digested group, LysC/Trypsin was introduced for a 1 hr pre-digestion prior to a 1 hr pepsin digestion.

### Liquid chromatography tandem mass spectrometry

The peptides of the digested commercial antibodies were analyzed on Q Exactive HF-X mass spectrometer (Thermo Scientific) connected to an EASY-nLC 1200 ultra-high-performance liquid chromatography system (Thermo Scientific). The peptides were loaded onto an Acclaim PepMap 100 (75µm x 2 cm) C18 (3 µm, 100 Å) pre-column and separated on an EASY-Spray column (Thermo Scientific; ID 75µm x 50 cm, column temperature 45°C) operated at a constant pressure of 800 bar. A linear gradient from 3 to 38% of 80% acetonitrile in aqueous 0.1% formic acid was run for 120 min at a flow rate of 350 nl min<sup>-1</sup>. One full MS scan (resolution 120 000 @ 200 m/z; mass range 350–1650 m/z) was followed by MS/MS scans (resolution 15000 @ 200 m/z) of the 15 most abundant ion signals. The isolation width window for the precursor ions was 1.3 m/z, they were fragmented using higher-energy collisional-induced dissociation (HCD) at a normalized collision energy of 28. Charge state screening was enabled, and precursor ions with unknown charge states and a charge state of 1, and over 6 were rejected. Data was additionally collected for non-tryptic digestions as above, but including peptides with a charge state of 1. The dynamic exclusion window was 10 s. The automatic gain control was set to 3e6 and 1e5 for MS and MS/MS with ion accumulation times of 45 ms and 30 ms, respectively.

### Computational analyses

**Spectra annotation.** A snakemake [53] was created for the DDA search. All DDA raw files were initially converted to Mascot generic format (MGF) by ThermoRawFileParser software. Ursgal package [54,55] was used as an interface for searching the spectra against data's



Uniprot reference proteome using five search engines, namely MSGFPlus [56] (version 2019.07.03), MS Amanda [57] (version 2.0.0.17442), Comet [58–60] (version 2019.01.rev5), X! Tandem [60] (version alanine), and OMSSA [61] (version 2.1.9). Optional Met Oxidation (UniMod: 35), along with the fixed Cys carbamidomethylation (UniMod: 4) modifications, were considered in this study. Individual engine results were validated by percolator [62] (version 3.4.0), while the Combine FDR algorithm was implemented for combining results from all search engines [63]. Moreover, a threshold of 1% peptide FDR was set for decisive candidate inclusion.

**De novo model generation and evaluation.** The process of creating a model involves 3 steps, namely: 1) establishing the training, validation, and test sets; 2) creation of the input files for DeepNovoV2 [42]; and 3) model training. Starting from the DDA search results over the Detroit 562 system, we extracted an equal number of annotated scans from each of the five protease datasets to make a suitable quantitative comparison of the resulting models. In that sense, limited by the dataset with the lowest number of unique peptides (gluc-dataset), we randomly selected 21492 peptides. These were then randomly divided into training, validation, and test sets in proportions of 90%, 5%, and 5%, respectively. For the second step, a snakemake workflow was created for the extraction of the selected spectra and generation of the features and MGF files. Finally, model training was done in 20 epochs [27,42]. For the feature extraction process, we consider a total of 12 ion-types, namely: a, b, y, a(2+), b(2+), y(2+), a-H<sub>2</sub>O, b-H<sub>2</sub>O, y-H<sub>2</sub>O, a-NH<sub>3</sub>, b-NH<sub>3</sub>, and y-NH<sub>3</sub>. Maximum peptide length and mass were adjusted to 4000 Da and 30, respectively. These models were called SEMs.

The evaluation of the initial models was accomplished through full cross-validation. This was done with the aim of obtaining a perspective on the performance of each test set, as well as overall. The same modifications employed in database search were considered for all of the *de novo* searches in this research. Additionally, the maximum deviation of the precursor mass was adjusted to 15ppm. Peptide recall was used as a measure of the quality of the models.

**Structural modeling.** The Fc and Fab domains of each antibody were *de novo* modeled separately by AlphaFold2 [64,65], considering MMseqs2 [66] to generate the multiple sequence alignment and homo-oligomer state of 1:1. For each selected model, the sidechains and the disulfide bridges were adjusted and relaxed using Rosetta relax protocol [67]. The loops in the hinge region were then re-modeled and characterized using DaReUS-Loop web server [68]. Finally, the full-length structure was relaxed, and all disulfide bridges (specifically in the hinge region) were adjusted using the Rosetta relax protocol. Visualization of the monoclonal antibodies was done through USCF Chimera software [69].

## Supporting information

**S1 Text. Singel Enzyme (SEM) and Multi Enzyme (MEM) models performance on the Giga test set. A.** Comparing all SEM models against the Kilo MEM model; **B.** Generalizability as a function of the number of SEM-combined training sets. These results show the same trends as the ones obtained on the Detroit test set.

(DOCX)

**S2 Text. De novo performance summary from the commercial monoclonal antibodies**

**(mAbs) samples. A.** Number of matching peptides extracted from each sample type across all models; **B.** Number of extracted peptides for each deep learning model; **C.** Length distribution for the mAbs matching peptides across all considered models. The color scheme for the samples and models is at the bottom.

(DOCX)

**S3 Text. Positional mAb data for the Giga+ MEM model *de novo* results.** **A.** Positional frequency values across all residues composing the ten mAb polypeptides proteins; **B.** Positional frequency information for the heavy and light subunits; **C.** Confident positional score (CS) values for the variable and constant domains of the heavy and light chains. For the color scheme, we used dark and light brown colors for the heavy and light chains, respectively. (DOCX)

**S1 Table. Sheets: ModelAttributes. GlobalRecall. LocalRecall. AntibodyCoverage.** (XLSX)

## Author Contributions

**Conceptualization:** Carlos Gueto-Tettay, Johan Malmström, Lars Malmström.

**Data curation:** Carlos Gueto-Tettay.

**Formal analysis:** Carlos Gueto-Tettay, Hamed Khakzad.

**Funding acquisition:** Johan Malmström, Lars Malmström.

**Investigation:** Carlos Gueto-Tettay, Di Tang, Lotta Happonen, Lars Malmström.

**Methodology:** Carlos Gueto-Tettay, Lars Malmström.

**Project administration:** Carlos Gueto-Tettay, Johan Malmström, Lars Malmström.

**Resources:** Di Tang, Lotta Happonen, Moritz Heusel, Johan Malmström, Lars Malmström.

**Software:** Carlos Gueto-Tettay, Lars Malmström.

**Supervision:** Johan Malmström, Lars Malmström.

**Visualization:** Carlos Gueto-Tettay.

**Writing – original draft:** Carlos Gueto-Tettay, Di Tang, Lotta Happonen, Hamed Khakzad, Johan Malmström, Lars Malmström.

**Writing – review & editing:** Carlos Gueto-Tettay, Lars Malmström.

## References

1. Han X, Aslanian A, Yates JR. Mass spectrometry for proteomics. *Curr Opin Chem Biol.* 2008; 12: 483–490. <https://doi.org/10.1016/j.cbpa.2008.07.024> PMID: 18718552
2. Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature.* 2016; 537: 347–355. <https://doi.org/10.1038/nature19949> PMID: 27629641
3. Schaffer LV, Millikin RJ, Miller RM, Anderson LC, Fellers RT, Ge Y, et al. Identification and quantification of proteoforms by mass spectrometry. *Proteomics.* 2019; 19: e1800361. <https://doi.org/10.1002/pmic.201800361> PMID: 31050378
4. Palm F, Chowdhury S, Wettemark S, Malmström J, Happonen L, Shannon O. Distinct Serotypes of Streptococcal M Proteins Mediate Fibrinogen-Dependent Platelet Activation and Proinflammatory Effects. *Infect Immun.* 2022; 90: e0046221. <https://doi.org/10.1128/IAI.00462-21> PMID: 34898252
5. Chowdhury S, Khakzad H, Bergdahl GE, Lood R, Ekstrom S, Linke D, et al. Streptococcus pyogenes Forms Serotype- and Local Environment-Dependent Interspecies Protein Complexes. *mSystems.* 2021; 6: e0027121. <https://doi.org/10.1128/mSystems.00271-21> PMID: 34581598
6. Gómez Toledo A, Sorrentino JT, Sandoval DR, Malmström J, Lewis NE, Esko JD. A systems view of the heparan sulfate interactome. *J Histochem Cytochem.* 2021; 69: 105–119. <https://doi.org/10.1369/0022155420988661> PMID: 33494649
7. Bahnan W, Wrighton S, Sundwall M, Bläckberg A, Larsson O, Höglund U, et al. Spike-Dependent Opsonization Indicates Both Dose-Dependent Inhibition of Phagocytosis and That Non-Neutralizing Antibodies Can Confer Protection to SARS-CoV-2. *Front Immunol.* 2021; 12: 808932. <https://doi.org/10.3389/fimmu.2021.808932> PMID: 35095897

8. Khakzad H, Happonen L, Tran Van Nhieu G, Malmström J, Malmström L. In vivo Cross-Linking MS of the Complement System MAC Assembled on Live Gram-Positive Bacteria. *Front Genet.* 2020; 11: 612475. <https://doi.org/10.3389/fgene.2020.612475> PMID: 33488677
9. Happonen L, Hauri S, Svensson Birkedal G, Karlsson C, de Neergaard T, Khakzad H, et al. A quantitative Streptococcus pyogenes-human protein-protein interaction map reveals localization of opsonizing antibodies. *Nat Commun.* 2019; 10: 2727. <https://doi.org/10.1038/s41467-019-10583-5> PMID: 31227708
10. Khakzad H, Happonen L, Karami Y, Chowdhury S, Bergdahl GE, Nilges M, et al. Structural determination of Streptococcus pyogenes M1 protein interactions with human immunoglobulin G using integrative structural biology. *PLoS Comput Biol.* 2021; 17: e1008169. <https://doi.org/10.1371/journal.pcbi.1008169> PMID: 33411763
11. Chowdhury S, Happonen L, Khakzad H, Malmström L, Malmström J. Structural proteomics, electron cryo-microscopy and structural modeling approaches in bacteria-human protein interactions. *Med Microbiol Immunol.* 2020; 209: 265–275. <https://doi.org/10.1007/s00430-020-00663-5> PMID: 32072248
12. Hauri S, Khakzad H, Happonen L, Teleman J, Malmström J, Malmström L. Rapid determination of quaternary protein structures in complex biological samples. *Nat Commun.* 2019; 10: 192. <https://doi.org/10.1038/s41467-018-07986-1> PMID: 30643114
13. Sjöholm K, Kilsgård O, Teleman J, Happonen L, Malmström L, Malmström J. Targeted Proteomics and Absolute Protein Quantification for the Construction of a Stoichiometric Host-Pathogen Surface Density Model. *Mol Cell Proteomics.* 2017; 16: S29–S41. <https://doi.org/10.1074/mcp.M116.063966> PMID: 28183813
14. Bundgaard L, Åhrman E, Malmström J, Auf dem Keller U, Walters M, Jacobsen S. Effective protein extraction combined with data independent acquisition analysis reveals a comprehensive and quantifiable insight into the proteomes of articular cartilage and subchondral bone. *Osteoarthr Cartil.* 2022; 30: 137–146. <https://doi.org/10.1016/j.joca.2021.09.006> PMID: 34547431
15. Bakochi A, Mohanty T, Pyl PT, Gueto-Tettay CA, Malmström L, Linder A, et al. Cerebrospinal fluid proteome maps detect pathogen-specific host response patterns in meningitis. *eLife.* 2021; 10. <https://doi.org/10.7554/eLife.64159> PMID: 33821792
16. Fisher J, Mohanty T, Karlsson CAQ, Khademi SMH, Malmström E, Frigyesi A, et al. Proteome Profiling of Recombinant DNase Therapy in Reducing NETs and Aiding Recovery in COVID-19 Patients. *Mol Cell Proteomics.* 2021; 20: 100113. <https://doi.org/10.1016/j.mcpro.2021.100113> PMID: 34139362
17. Wilk L, Happonen L, Malmström J, Herwald H. Comprehensive Mass Spectrometric Survey of Streptococcus pyogenes Subcellular Proteomes. *J Proteome Res.* 2018; 17: 600–617. <https://doi.org/10.1021/acs.jproteome.7b00701> PMID: 29160079
18. Malmström E, Kilsgård O, Hauri S, Smeds E, Herwald H, Malmström L, et al. Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics. *Nat Commun.* 2016; 7: 10261. <https://doi.org/10.1038/ncomms10261> PMID: 26732734
19. Olsen JV, Ong S-E, Mann M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics.* 2004; 3: 608–614. <https://doi.org/10.1074/mcp.T400003-MCP200> PMID: 15034119
20. Lakbub JC, Shipman JT, Desaire H. Recent mass spectrometry-based techniques and considerations for disulfide bond characterization in proteins. *Anal Bioanal Chem.* 2018; 410: 2467–2484. <https://doi.org/10.1007/s00216-017-0772-1> PMID: 29256076
21. Winther JR, Thorpe C. Quantification of thiols and disulfides. *Biochim Biophys Acta.* 2014; 1840: 838–846. <https://doi.org/10.1016/j.bbagen.2013.03.031> PMID: 23567800
22. Giansanti P, Tsiatsiani L, Low TY, Heck AJR. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc.* 2016; 11: 993–1006. <https://doi.org/10.1038/nprot.2016.057> PMID: 27123950
23. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res.* 2010; 9: 1323–1329. <https://doi.org/10.1021/pr900863u> PMID: 20113005
24. Yang H, Li Y-C, Zhao M-Z, Wu F-L, Wang X, Xiao W-D, et al. Precision De Novo Peptide Sequencing Using Mirror Proteases of Ac-LysargiNase and Trypsin for Large-scale Proteomics. *Mol Cell Proteomics.* 2019; 18: 773–785. <https://doi.org/10.1074/mcp.TIR118.000918> PMID: 30622160
25. Samodova D, Hosfield CM, Cramer CN, Giuli MV, Cappellini E, Franciosa G, et al. Proalanase is an effective alternative to trypsin for proteomics applications and disulfide bond mapping. *Mol Cell Proteomics.* 2020; 19: 2139–2157. <https://doi.org/10.1074/mcp.TIR120.002129> PMID: 33020190
26. Vitorino R, Guedes S, Trindade F, Correia I, Moura G, Carvalho P, et al. De novo sequencing of proteins by mass spectrometry. *Expert Rev Proteomics.* 2020; 17: 595–607. <https://doi.org/10.1080/14789450.2020.1831387> PMID: 33016158

27. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci USA*. 2017; 114: 8247–8252. <https://doi.org/10.1073/pnas.1705691114> PMID: 28720701
28. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*. 2003; 17: 2337–2342. <https://doi.org/10.1002/rcm.1196> PMID: 14558135
29. Fischer B, Roth V, Roos F, Grossmann J, Baginsky S, Widmayer P, et al. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem*. 2005; 77: 7265–7273. <https://doi.org/10.1021/ac0508853> PMID: 16285674
30. Muth T, Renard BY. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinformatics*. 2018; 19: 954–970. <https://doi.org/10.1093/bib/bbx033> PMID: 28369237
31. Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol*. 1999; 6: 327–342. <https://doi.org/10.1089/106652799318300> PMID: 10582570
32. Chi H, Chen H, He K, Wu L, Yang B, Sun R-X, et al. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J Proteome Res*. 2013; 12: 615–625. <https://doi.org/10.1021/pr3006843> PMID: 23272783
33. Yang H, Chi H, Zeng W-F, Zhou W-J, He S-M. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*. 2019; 35: i183–i190. <https://doi.org/10.1093/bioinformatics/btz366> PMID: 31510687
34. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem*. 2005; 77: 964–973. <https://doi.org/10.1021/ac048788h> PMID: 15858974
35. Wen B, Zeng W-F, Liao Y, Shi Z, Savage SR, Jiang W, et al. Deep learning in proteomics. *Proteomics*. 2020; 20: e1900335. <https://doi.org/10.1002/pmic.201900335> PMID: 32939979
36. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521: 436–444. <https://doi.org/10.1038/nature14539> PMID: 26017442
37. Tsuruoka Y. [deep learning and natural language processing]. *Brain Nerve*. 2019; 71: 45–55. <https://doi.org/10.11477/mf.1416201215> PMID: 30630129
38. Kriegeskorte N, Golan T. Neural network models and deep learning. *Curr Biol*. 2019; 29: R231–R236. <https://doi.org/10.1016/j.cub.2019.02.034> PMID: 30939301
39. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 2021; 8: 53. <https://doi.org/10.1186/s40537-021-00444-8> PMID: 33816053
40. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: A systematic review. *IEEE Access*. 2019; 7: 19143–19165. <https://doi.org/10.1109/ACCESS.2019.2896880>
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. *arXiv*. 2017. <https://doi.org/10.48550/arxiv.1706.03762>
42. Qiao R, Tran NH, Xin L, Shan B, Li M, Ghodsi A. DeepNovoV2: Better de novo peptide sequencing with deep learning. *arXiv*. 2019. <https://doi.org/10.48550/arxiv.1904.08514>
43. Guthals A, Gan Y, Murray L, Chen Y, Stinson J, Nakamura G, et al. De novo MS/MS sequencing of native human antibodies. *J Proteome Res*. 2017; 16: 45–54. <https://doi.org/10.1021/acs.jproteome.6b00608> PMID: 27779884
44. Chen J, Zheng Q, Hammers CM, Ellebrecht CT, Mukherjee EM, Tang H-Y, et al. Proteomic Analysis of Pemphigus Autoantibodies Indicates a Larger, More Diverse, and More Dynamic Repertoire than Determined by B Cell Genetics. *Cell Rep*. 2017; 18: 237–247. <https://doi.org/10.1016/j.celrep.2016.12.013> PMID: 28052253
45. Boutz DR, Horton AP, Wine Y, Lavinder JJ, Georgiou G, Marcotte EM. Proteomic identification of monoclonal antibodies from serum. *Anal Chem*. 2014; 86: 4758–4766. <https://doi.org/10.1021/ac4037679> PMID: 24684310
46. Diedrich JK, Pinto AFM, Yates JR. Energy dependence of HCD on peptide fragmentation: stepped collisional energy finds the sweet spot. *J Am Soc Mass Spectrom*. 2013; 24: 1690–1699. <https://doi.org/10.1007/s13361-013-0709-7> PMID: 23963813
47. Cox KA, Gaskell SJ, Morris M, Whiting A. Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions. *J Am Soc Mass Spectrom*. 1996; 7: 522–531. [https://doi.org/10.1016/1044-0305\(96\)00019-0](https://doi.org/10.1016/1044-0305(96)00019-0) PMID: 24203424
48. Tabb DL, Huang Y, Wysocki VH, Yates JR. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem*. 2004; 76: 1243–1248. <https://doi.org/10.1021/ac0351163> PMID: 14987077

49. van Dongen WD, Ruijters HF, Luinge HJ, Heerma W, Haverkamp J. Statistical analysis of mass spectral data obtained from singly protonated peptides under high-energy collision-induced dissociation conditions. *J Mass Spectrom.* 1996; 31: 1156–1162. [https://doi.org/10.1002/\(SICI\)1096-9888\(199610\)31:10<1156::AID-JMS407>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1096-9888(199610)31:10<1156::AID-JMS407>3.0.CO;2-T) PMID: 8916424
50. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA. The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol.* 2007; 25: 21–50. <https://doi.org/10.1146/annurev.immunol.25.022106.141702> PMID: 17029568
51. Zauner G, Selman MHJ, Bondt A, Rombouts Y, Blank D, Deelder AM, et al. Glycoproteomic analysis of antibodies. *Mol Cell Proteomics.* 2013; 12: 856–865. <https://doi.org/10.1074/mcp.R112.026005> PMID: 23325769
52. Castilho A, Gruber C, Thader A, Oostenbrink C, Pechlaner M, Steinkellner H, et al. Processing of complex N-glycans in IgG Fc-region is affected by core fucosylation. *MAbs.* 2015; 7: 863–870. <https://doi.org/10.1080/19420862.2015.1053683> PMID: 26067753
53. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* 2012; 28: 2520–2522. <https://doi.org/10.1093/bioinformatics/bts480> PMID: 22908215
54. Kremer LPM, Leufken J, Oyunchimeg P, Schulze S, Fufezan C, Ursгал, Universal Python Module Combining Common Bottom-Up Proteomics Tools for Large-Scale Analysis. *J Proteome Res.* 2016; 15: 788–794. <https://doi.org/10.1021/acs.jproteome.5b00860> PMID: 26709623
55. Schulze S, Igiraneza AB, Kösters M, Leufken J, Leidel SA, Garcia BA, et al. Enhancing open modification searches via a combined approach facilitated by ursгал. *J Proteome Res.* 2021; 20: 1986–1996. <https://doi.org/10.1021/acs.jproteome.0c00799> PMID: 33514075
56. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun.* 2014; 5: 5277. <https://doi.org/10.1038/ncomms6277> PMID: 25358478
57. Dorfer V, Pichler P, Stranzl T, Stadlmann J, Taus T, Winkler S, et al. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J Proteome Res.* 2014; 13: 3679–3684. <https://doi.org/10.1021/pr500202e> PMID: 24909410
58. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics.* 2013; 13: 22–24. <https://doi.org/10.1002/pmic.201200439> PMID: 23148064
59. Eng JK, Hoopmann MR, Jahan TA, Egertson JD, Noble WS, MacCoss MJ. A deeper look into Comet—implementation and features. *J Am Soc Mass Spectrom.* 2015; 26: 1865–1874. <https://doi.org/10.1007/s13361-015-1179-x> PMID: 26115965
60. Schweppe DK, Eng JK, Yu Q, Bailey D, Rad R, Navarrete-Perea J, et al. Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J Proteome Res.* 2020; 19: 2026–2034. <https://doi.org/10.1021/acs.jproteome.9b00860> PMID: 32126768
61. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. *J Proteome Res.* 2004; 3: 958–964. <https://doi.org/10.1021/pr0499491> PMID: 15473683
62. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom.* 2016; 27: 1719–1727. <https://doi.org/10.1007/s13361-016-1460-7> PMID: 27572102
63. Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics.* 2009; 9: 1220–1229. <https://doi.org/10.1002/pmic.200800473> PMID: 19253293
64. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021; 596: 583–589. <https://doi.org/10.1038/s41586-021-03819-2> PMID: 34265844
65. Mirdita M, Ovchinnikov S, Steinegger M. ColabFold—Making protein folding accessible to all. *BioRxiv.* 2021. <https://doi.org/10.1101/2021.08.15.456425>
66. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol.* 2017; 35: 1026–1028. <https://doi.org/10.1038/nbt.3988> PMID: 29035372
67. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods.* 2020; 17: 665–680. <https://doi.org/10.1038/s41592-020-0848-2> PMID: 32483333
68. Karami Y, Rey J, Postic G, Murail S, Tufféry P, de Vries SJ. DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Res.* 2019; 47: W423–W428. <https://doi.org/10.1093/nar/gkz403> PMID: 31114872

69. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem.* 2004; 25: 1605–1612. <https://doi.org/10.1002/jcc.20084> PMID: [15264254](https://pubmed.ncbi.nlm.nih.gov/15264254/)