






# Comprehensive evaluation of peptide *de novo* sequencing tools for monoclonal antibody assembly

Denis Beslic , Georg Tscheuschner , Bernhard Y. Renard , Michael G. Weller  and Thilo Muth 

Corresponding authors. D. Beslic, Robert Koch Institute, ZKI-PH 3, Nordufer 20, 13353 Berlin, Germany. E-mail: beslicd@rki.de; G. Tscheuschner, Federal Institute for Materials Research and Testing (BAM), Richard-Willstätter-Straße 11, 12489 Berlin, Germany. E-mail: georg.tscheuschner@bam.de; B.Y. Renard, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany. E-mail: bernhard.renard@hpi.de; M.G. Weller, Federal Institute for Materials Research and Testing (BAM), Richard-Willstätter-Straße 11, 12489 Berlin, Germany. E-mail: michael.weller@bam.de; T. Muth, Federal Institute for Materials Research and Testing (BAM), Unter den Eichen 87, 12205 Berlin, Germany. E-mail: thilo.muth@bam.de

## Abstract

Monoclonal antibodies are biotechnologically produced proteins with various applications in research, therapeutics and diagnostics. Their ability to recognize and bind to specific molecule structures makes them essential research tools and therapeutic agents. Sequence information of antibodies is helpful for understanding antibody–antigen interactions and ensuring their affinity and specificity. *De novo* protein sequencing based on mass spectrometry is a valuable method to obtain the amino acid sequence of peptides and proteins without a priori knowledge. In this study, we evaluated six recently developed *de novo* peptide sequencing algorithms (Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo), which were not specifically designed for antibody data. We validated their ability to identify and assemble antibody sequences on three multi-enzymatic data sets. The deep learning-based tools Casanovo and PointNovo showed an increased peptide recall across different enzymes and data sets compared with spectrum-graph-based approaches. We evaluated different error types of *de novo* peptide sequencing tools and their performance for different numbers of missing cleavage sites, noisy spectra and peptides of various lengths. We achieved a sequence coverage of 97.69–99.53% on the light chains of three different antibody data sets using the de Bruijn assembler ALPS and the predictions from Casanovo. However, low sequence coverage and accuracy on the heavy chains demonstrate that complete *de novo* protein sequencing remains a challenging issue in proteomics that requires improved *de novo* error correction, alternative digestion strategies and hybrid approaches such as homology search to achieve high accuracy on long protein sequences.

**Keywords:** *de novo* peptide sequencing, bioinformatics, benchmarking study, monoclonal antibodies, mass spectrometry

## Introduction

Monoclonal antibodies (mAbs) are immunoglobulins of unique specificity generated artificially in laboratories to mimic antibodies produced by the immune system [1]. Their reproducibility under certain conditions and high binding affinity to target molecules make them essential to various diagnostic and analytical applications in immunology, clinical chemistry, food chemistry, environmental analysis, biochemistry, therapeutics and medicine [2–4]. Recently, multiple authors reported how antibodies lack proper classification and identification as research tools, thereby causing a so-called reproducibility crisis [5]. The results of multiple landmark papers could not be replicated because mAbs often lacked crucial quality control steps for correct characterization [6, 7]. One essential step for improving the research quality includes the confirmation of the amino acid sequence [8, 9]. In addition, retrieving sequence information of antibodies is crucial for understanding the structural basis of antibody–antigen binding, recognition and interaction [10]. The structural basis for the specificity in protein–protein interactions lies in the sequence diversity of antibodies. The majority of sequence diversity focuses on the hypervariable loops within

the variable regions of antibodies, called complementarity-determining regions (CDRs), which are mainly responsible for the interaction between the antibody and their target structures [10, 11]. Most established methods for antibody *de novo* sequencing rely on sequencing mRNA from hybridoma cells. However, these approaches all depend on the availability of pure clones of antibody-producing cells [12]. Moreover, crucial posttranslational modifications, which affect antigen binding, developability and effector functions, cannot be detected by DNA sequencing [3]. Hence, approaches to sequence the antibody on protein level are necessary.

Tandem mass spectrometry (MS/MS) is a powerful method for retrieving the amino acid sequence of peptides. Typically, in standard shotgun proteomics, protein samples are digested with proteolytic enzymes into shorter peptides, which are more suitable for analysis by MS/MS [13]. To obtain sequential information from novel or unknown proteins, *de novo* peptide sequencing is commonly used, which identifies peptides directly from MS/MS spectra without relying on a sequence database [14]. Here, each amino acid is derived by computing mass differences of ions from a fragmented peptide. As the manual characterization

**Denis Beslic** is a doctoral candidate at the Robert Koch Institute in Berlin, Germany.

**Georg Tscheuschner** is a doctoral candidate at the Federal Institute for Materials Research and Testing (BAM) in Berlin, Germany.

**Bernhard Y. Renard** is professor for data analytics and computational statistics at Hasso-Plattner-Institute and the University of Potsdam in Potsdam, Germany.

**Michael G. Weller** is the head of section 1.5 (Protein Analysis) at the Federal Institute for Materials Research and Testing (BAM) in Berlin, Germany.

**Thilo Muth** is the head of section S.3 (eScience) at the Federal Institute for Materials Research and Testing (BAM) in Berlin, Germany.

**Received:** September 9, 2022. **Revised:** October 25, 2022. **Accepted:** November 10, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

of peptides using *de novo* sequencing can be very time consuming and challenging, a variety of algorithms have been developed to differentiate signal ion peaks from noise peaks to predict the correct peptide sequence [14–16]. Recent advances in deep learning (DL) have marked an important milestone for database-independent prediction of peptide sequences from MS/MS data [16]. The encoder–decoder architecture was designed to solve specific tasks in sequence-to-sequence learning [17]. Tran *et al.* [18] employed convolutional neural networks (CNNs) to encode mass spectra while using recurrent neural networks (RNNs) as a decoder to predict the amino acids of peptide sequences one by one. Their method DeepNovo outperformed state-of-the-art methods at that time. Multiple methods have been published based on the network architecture of DeepNovo, namely, DeepNovo-DIA [19], SMSNet [20] and PointNovo [21]. More recently, the transformer-based framework Casanovo showed promising results for the prediction of peptide sequences [22].

Although peptide *de novo* sequencing has improved in recent years, the full-length assembly of protein sequences poses another challenging task. In most cases, database search algorithms, such as MSGF+ [23], infer the correct proteins from identified peptide sequences [24]. However, the determination of protein sequences, which are not part of public databases, limits the feasibility of this approach. In the case of unknown antibodies, the variable sequence is not available and cannot be derived from database search algorithms [25]. Hence, *de novo* peptide sequencing and the assembly of the predicted peptides are necessary for assessing the amino acid sequence of unknown antibodies. Currently, only a few developed methods were reported for database-independent full-length antibody *de novo* sequencing and assembly, for instance, meta-SPS [26], ALPS [27], pTa [25] and MuCS [28]. Meta-SPS utilized overlapping fragment ion peaks from different spectra to construct meta-contigs before *de novo* sequencing. Across six diverse proteins and the aBTLA antibody, the authors observed a sequence coverage between 68 and 99%. Nonetheless, meta-SPS faces multiple limitations and is not combinable with recently developed *de novo* sequencing algorithms [26]. Tran *et al.* analyzed antibodies using PEAKS *de novo* [29], PEAKS DB [30] and the homology software SPIDER [31] in a complementary way. The results from these three algorithms serve as input for their *de Bruijn* assembler ALPS. Still, despite using homology and database search algorithms, the authors inspected a fragmented and incomplete assembly of long antibody sequences, particularly at the variable region of the heavy chain [27]. Thus, *de novo* sequencing of proteins remains a challenging and important problem to date.

Most publications regarding new *de novo* peptide sequencing approaches include a performance comparison of recently developed tools [14, 32, 33], yet, to our knowledge, there is no published independent evaluation of different *de novo* sequencing algorithms on antibody data sets. Moreover, newly developed *de novo* peptide sequencing tools rarely include antibodies for benchmarking their method in comparison to already existing tools [20–22]. Only the authors of DeepNovo used antibodies as an example application for their tool [18]. *De novo* sequencing studies involving antibody data mostly deal with validating antibody-specific assembly tools [25–27] or introducing alternative experimental methods [12, 34]. In this study, we present a performance evaluation of six recently developed *de novo* sequencing algorithms (Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo), which we chose based on their availability and performance in previous studies [14, 21, 22, 33]. In contrast to previous studies, we evaluate *de novo* peptide sequencing tools on various

enzymatic antibody data sets. Furthermore, we investigate common error types, the impact of noisy spectra and missing fragmentation ions. To compare the ability of previously mentioned tools to reconstruct full-length protein sequences without additional database algorithms, we employed the *de Bruijn* assembler ALPS. Finally, we discuss possible solutions and the demanding challenges of *de novo* antibody sequencing.

## Materials and methods

### Antibody data sets

We analyzed data from three publicly available antibody data sets. The data sets were downloaded from the proteomics MS data repositories PRIDE [35] and MASSIVE [36]. Table 1 gives an overview of the three evaluated antibody data sets, showing the reference, available digestion, mass instrument, ionization type and fragment ion resolution. The Glu-C file of IgG1-Human-LC and the trypsin file of WIGG1-Mouse-LC were corrupted and therefore not included in our analysis. In total, we analyzed 25 different MS/MS experiments (Supplementary Figure S1).

### Data processing

#### *De novo* sequencing methods

Table 2 provides an overview of the evaluated *de novo* peptide sequencing algorithms. Furthermore, it provides information about the algorithmic paradigm, the project website and the corresponding reference.

#### Description of *de novo* sequencing algorithms used

Novor [37] is based on a decision tree scoring function to select peptide predictions. pNovo 3 [38] employs a learning-to-rank framework using gap features and predictions from pDeep [39] to improve the scoring of peptide sequences. Novor and pNovo 3 were developed based on the spectrum-graph approach while using extensive machine-learning algorithms for an enhanced scoring function. DeepNovo [18] was the first approach to incorporate the encoder–decoder paradigm for *de novo* peptide sequencing. SMSNet [20] uses a similar CNN- and RNN-based framework but additionally includes optional post-processing and a shift layer in the encoder module. PointNovo [21] adopts an order-invariant network structure for the prediction process of higher-resolution data. Casanovo [22] employs a transformer-based framework to process and predict sequences of amino acids instead of using RNNs.

#### Preprocessing

Each instrument vendor uses its own file formats to store results from MS/MS experiments. These raw files need to be converted to open-format files to be compatible with *de novo* sequencing tools [40]. We reformatted the raw MS/MS data files from the previously mentioned data sets to Mascot Generic Format (MGF) using ProteoWizard [41]. A MGF file stores the *m/z* and intensity pairs of multiple mass spectra in a single text format. *De novo* sequencing tools predict amino acid residues by accessing the mass differences between the MS/MS peaks [14].

#### Parameters for *de novo* sequencing algorithms

We executed Novor (v.1.05) via the DeNovoGUI command-line interface (v.1.16.6) [42]. We ran pNovo 3 (v.3.1.3) via its executable GUI, which included pre-trained models for specific enzymes. To perform a fair comparison between spectrum-graph-based tools like Novor and pNovo 3, which are released only with pre-trained models, and the DL algorithms, DeepNovo (v.PNAS), SMSNet,

**Table 1.** Overview of evaluated antibody data sets. For each data set, we provided the name of the data set, the ID, the mass instrument, the ionization type, the fragment ion resolution in FWHM, a reference and the number of proteolytic enzymes in the data set

Data set name	Database ID	Mass instrument	Ionization type	Resolution	Ref.	Enzymes
IgG1-Human	MSV000079801	LTQ Orbitrap	HCD	17 500	[27]	Trypsin, chymotrypsin, asp-N, lys-C, glu-C, proteinase K
WlgG1-Mouse	MSV000079801	LTQ Orbitrap	HCD	17 500	[27]	Trypsin, asp-N, chymotrypsin
Herceptin	PXD023419	Orbitrap Fusion	Stepped HCD and EThcD	30 000	[12]	Trypsin, thermolysin, lys-N, lys-C, glu-C, asp-N, aLP, chymotrypsin, elastase

**Table 2.** Overview of all *de novo* sequencing tools used in this study. For each algorithm, the name of the tool, the algorithmic paradigm, the year of the publication, the reference and the project website of the corresponding method are displayed

Software	Algorithmic paradigm	Year	Ref.	Project website
Novor	Spectrum graph, machine learning, decision tree	2015	[37]	<a href="http://rapidnovor.com/">rapidnovor.com/</a>
pNovo 3	Spectrum graph, machine learning, SVM	2019	[38]	<a href="http://i.pfind.org/">i.pfind.org/</a>
DeepNovo	DL, CNN + RNN	2017	[18]	<a href="https://github.com/nh2tran/DeepNovo/">github.com/nh2tran/DeepNovo/</a>
SMSNet	DL, CNN + RNN	2019	[20]	<a href="https://github.com/cmb-chula/SMSNet/">github.com/cmb-chula/SMSNet/</a>
PointNovo	DL, PointNet+RNN	2021	[21]	<a href="https://github.com/volpato30/PointNovo/">github.com/volpato30/PointNovo/</a>
Casanovo	DL, transformer	2022	[22]	<a href="https://github.com/Noble-Lab/casanovo">github.com/Noble-Lab/casanovo</a>

PointNovo (v.0.0.1) and Casanovo (v.3.0.0), we trained all DL-based tools on high-resolution MS/MS data from the human proteome using the HCD library from MassIVE, which consists of 1 114 503 different peptides [36]. Training them on specific antibody data would give DL programs an unfair advantage compared with pre-trained software. We split the spectra into training, validation and test sets at a ratio of 98:1:1 while making sure that the split data sets did not share any common peptides. Each model was trained for 10 epochs using pre-defined parameters from each tool. We executed all tools at a precursor tolerance of 10 ppm and fragment mass tolerance of 0.02 Da. For each algorithm, carbamidomethylation of cysteine (C + 57.02 Da) was set as a fixed modification. Oxidation of methionine (M + 15.99 Da) and deamidation of asparagine and glutamine (N + 0.98 Da and G + 0.98 Da) were set as variable modifications. DeNovoGUI and the DL tools were executed on a Linux server machine (100 cores, 64GB RAM). We executed pNovo 3 on a Windows 64-bit computer since the software was not supported by a Linux operating system.

### Assembly of identified peptides

The predicted peptides were further processed by the *de Bruijn* sequence assembler ALPS [27] to evaluate the ability of different *de novo* sequencing tools to reconstruct complete protein sequences. As described by the authors, a *k*-mer size of 7 ensures a sufficiently high coverage of the amino acid sequence while preventing repetitiveness of the resulting contigs at the same time. ALPS takes the *de novo* confidence score into consideration for the assembly, but it could generate incorrect results using a high amount of low-confidence *k*-mers. The authors of DeepNovo recommend removing sequence contaminants from *de novo* sequencing results by excluding peptides with a confidence score below 50 to improve the quality of the assembly [18]. Since every single *de novo* sequencing algorithm calculates its confidence score in a different manner, we chose the threshold for the confidence score based on the amino acid-level precision. We removed peptide sequences below the confidence score of each tool for

which the AA precision was below 50%. This aims to filter out low-quality predictions and, at the same time, ensures that we do not miss correctly predicted peptides, which have been assigned a low confidence score by the corresponding tool. We aligned the target contigs with the ground truth antibody sequence to classify the assembly results using BLAST [43]. Based on the alignments of the top contigs, we calculated the protein coverage and accuracy. The target sequence was regarded as being covered in case a contig was aligned to the target (sub)sequence. We calculated the accuracy by the number of correct sequence calls, which were aligned to the target sequence.

## Evaluation metrics

### Database search

For validating *de novo* sequencing algorithms, we compared each prediction to a pseudo-ground truth, which is commonly obtained by database search [14,44]. Since the evaluated data sets do not include a labeled ground truth for each spectrum, we performed a database search using the antibody sequences as our protein database. We used the combined results of the database algorithms MS-GF+ [23] and X!Tandem [45], which were both executed via SearchGUI (v.4.1.7) [46] and post-processed via PeptideShaker (v.2.2.2) [47] on a 64-bit Windows computer. We filtered all resulting peptide-spectrum matches (PSMs) using a false discovery rate (FDR) of 1%. The combined results of two database algorithms and an FDR rate of 1% would generate a reliable pseudo-ground truth for the evaluation of the *de novo* sequencing tools. We chose the cleavage parameters according to the enzyme used in the provided input file. Furthermore, the search parameters included the same modifications and mass tolerance that we selected for the *de novo* sequencing algorithms.

### Recall and precision

We compared the predictions of each *de novo* sequencing algorithm with the pseudo-ground truth peptides, which were identified by database search. Recall and accuracy were measured at

the peptide and amino acid level. The performance at the amino acid level was measured by matching amino acids between the prediction and the ground truth. We applied the same evaluation metric adapted by DeepNovo, Novor and PointNovo [18,21,37]: amino acids were considered as matched ones if their masses were different by  $<0.1$  Da and if the prefix masses before them were different by  $<0.5$  Da.

### Identification of fragment ions and noise

To evaluate the amount of noise and missing fragment ions in each spectrum, we labeled each peak as a peptide peak or a noise peak using the Pyteomics framework [48]. For each cleavage site, we tried to identify eight different ion types (b, b(2+), b-NH<sub>3</sub>, b-H<sub>2</sub>O, y, y(2+), y-NH<sub>3</sub> and y-H<sub>2</sub>O) since all evaluated *de novo* sequencing algorithms take these ion types into consideration for retrieving the peptide sequence. If possible, we matched these ion types to corresponding peaks in the spectra within a tolerance of 0.5 Da. Otherwise, we declared the cleavage site as missing. We only considered noise peaks if their intensity exceeded the median noise intensity for each data set. The number of noise peaks above this threshold was used to calculate the noise factor, which is defined as the ratio of the number of high-intensity peaks and the number of fragment ion peaks. McDonnell et al. [33] applied this approach recently in their evaluation of *de novo* sequencing algorithms.

## Results

### Performance of *de novo* sequencing algorithms on antibody data at the peptide and amino acid level

We evaluated six state-of-the-art *de novo* peptide sequencing algorithms, namely, Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo. For this purpose, we used the antibody data sets described in section Antibody Data sets to measure the accuracy across different enzymes using metrics specified in section Evaluation Metrics. Using three antibody data sets, we relied on 183 873 MS/MS scans, from which 23 844 peptides were identified with database search. Peptide identifications by database tools served as ground truth for evaluating the predictions from *de novo* sequencing tools. By comparing *de novo* sequencing results to this reference, we were able to identify the number of correctly predicted amino acids and peptides for each tool.

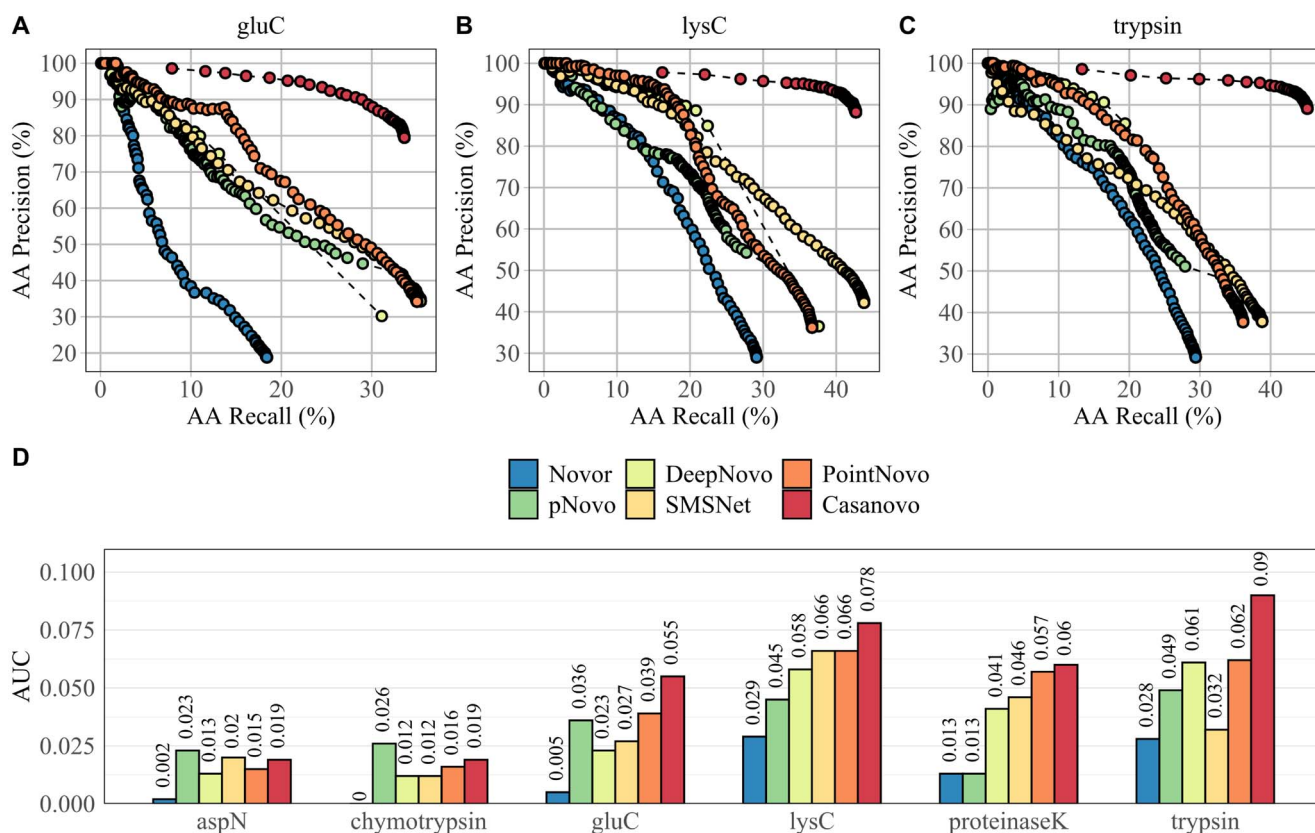
Each algorithm generates a confidence score along the predicted sequence to reflect its quality. Setting a threshold to the confidence score outputs different sets of predicted peptides. A high threshold would show a small number of peptides with high precision, but it would exclude a large part of the data set, consequently reducing the recall. Here, we used different thresholds of the confidence score to draw precision-recall (PR) curves and used the area under the curve (AUC) as a summary metric for the accuracy of *de novo* sequencing results. Figure 1 displays the PR curves (A–C) and the AUC (D) of *de novo* sequencing tools across six different enzymes of the IgG1-Human-HC data set. Casanovo shows the highest AUC value across four enzymatic data sets because of its high AA precision compared with all other algorithms. All six evaluated algorithms display an overall higher AUC on trypsin and lys-C than on other proteases. The performance is generally lower on the enzymatic data sets of asp-N and chymotrypsin. The lower efficiency of non-tryptic enzymes for the prediction of peptide sequences was reported in different publications and had several reasons [49–51]. First, trypsin shows a higher number of PSMs, which is caused by a bias of database search algorithms toward peptides digested with trypsin [52].

Furthermore, peptides digested with trypsin are better suited for HCD fragmentation since they include at least one positive charge at each terminus, generating reliable b- and y-ion fragmentation patterns. In contrast, non-tryptic proteases may lack positive-charged termini, which makes it more challenging to identify the correct peptide [53]. The AUC on asp-N, chymotrypsin, glu-C and proteinase-K is considerably lower across all tools because of their distinct cleavage patterns. The general low AUC across all enzymes can be explained by the differences between the training data and the evaluated antibodies. The DL-based tools were not trained on antibody-specific data but on peptides from the human proteome data, which were derived from various experimental conditions [36]. Moreover, the evaluated data sets include a smaller number of unique peptides compared with available benchmarking data sets used in other studies [18, 21, 22]. The high AUC of pNovo 3 on non-tryptic peptides can be attributed to the pre-trained models for specific enzymes. A pre-trained enzyme-specific model was not available for proteinase-K, which explains the low accuracy of pNovo 3 on this data set.

Figure 2 displays the total peptide recall (A), amino acid recall (B) and amino acid precision (C) across all six enzymatic cleavages of IgG1-Human-HC. In contrast to the results shown in Figure 1, we used all predictions from each tool regardless of their confidence score. Here, either SMSNet or PointNovo shows the highest amino acid recall in comparison to other tools on proteinase-K, asp-N, glu-C and lys-C (Figure 2B). Regarding the recall on peptide level, Casanovo exhibits the highest number of correct peptide predictions compared with all other *de novo* algorithms across all enzymes demonstrating the advantage of using transformers for peptide sequencing. Furthermore, Casanovo predicts amino acids with overall superior precision (Figure 2C). As Yilmaz et al. explained, the precursor *m/z* filter of Casanovo results in a prioritization of predicting full peptide sequences over partially correct subsequences [22]. Hence, Casanovo displays a very high AA precision and peptide recall, whereas its recall on amino acid level is comparable to PointNovo and SMSNet. The high accuracy of PointNovo can be attributed to its order-invariant networks, which have been applied for 3D recognition tasks and showed superior performance compared with state-of-the-art methods [54]. SMSNet profits from its shift layer, which helps to derive amino acids from MS/MS peaks [20]. Besides differences in the network structures, the number of fragment ions for predicting each amino acid residue plays another important role. While Novor and DeepNovo use eight ion types for predicting each position (y, y(2+), y-NH<sub>2</sub>, y-H<sub>2</sub>O, b, b(2+), b-NH<sub>2</sub> and b-H<sub>2</sub>O), SMSNet and pNovo 3 take nine ion types into consideration for inferring peptides from spectrum peaks. Moreover, PointNovo examines 12 ion types to calculate theoretical *m/z* values at each prediction step. In contrast, the transformer-based framework of Casanovo processes the entire set of spectrum peaks at once. pNovo 3 predicted a similar or even higher number of correct peptides of the non-tryptic data sets asp-N and chymotrypsin than the DL tools because of its enzyme-specific models. Furthermore, pNovo 3 shows high accuracy on amino acid level compared with PointNovo, SMSNet and DeepNovo (Figure 2C) because of its extensive reranking process. Nonetheless, pNovo 3 cannot predict amino acids with the same precision as Casanovo.

### Evaluation of error types

Following the performance on peptide and amino acid level, we evaluated the source of incorrect predictions of the different *de novo* sequencing algorithms. McDonnell et al. [33] reported previously that missing fragment ions and noise peaks pose a challenge



**Figure 1.** The PR curves of Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo for glu-C (A), lys-C (B) and trypsin (C) of the IgG1-Human-HC data set. The AUC of the six algorithms for each PR curve and each enzyme of IgG1-Human-HC (D).

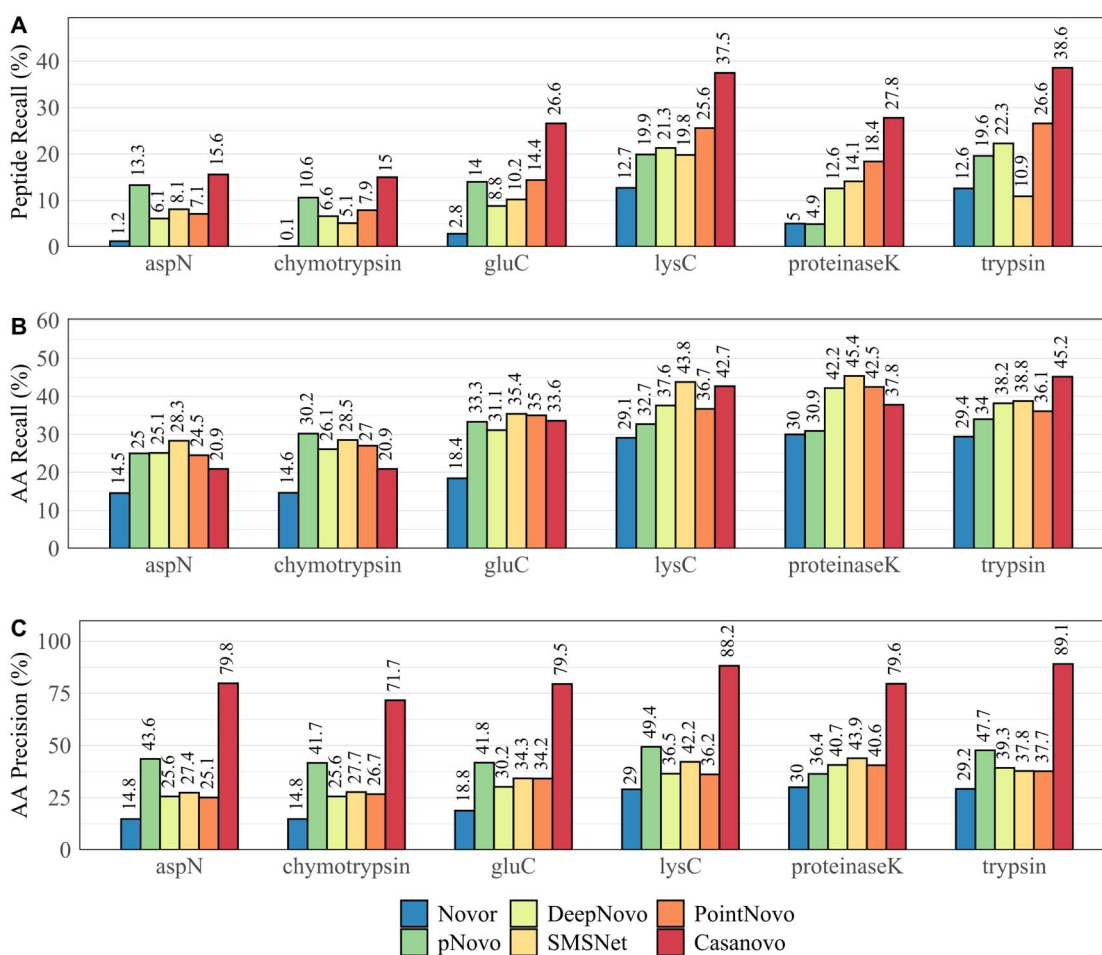
for *de novo* sequencing algorithms. We observed that 90.51% of all 23 227 validated spectra were missing at least one fragment ion. Furthermore, we detected that 84.32% of all peaks from these spectra were classified as noise peaks. In Figure 3, we show the peptide recall for different numbers of missing cleavage sites and different noise factors of all validated spectra. As expected, *de novo* sequencing algorithms tend to identify a higher number of correct peptides from spectra with a lower amount of missing fragment ions (Figure 3A). Missing fragment ions decrease the overall performance of all *de novo* sequencing tools. Nonetheless, Casanovo displays a superior performance on spectra with up to eight missing cleavage sites. On spectra with at least four missing cleavage sites, all other DL-based tools show a low peptide recall of 9.56–14.62%, whereas Casanovo performs considerably better. Novor shows a noticeably lower performance compared with all other algorithms.

When viewed alone, the noise factor of different spectra does not have a strong effect on the accuracy of the *de novo* sequencing algorithms (Figure 3B). As McDonnell stated, this is because of the stronger influence of the number of missing fragmentation sites on the peptide recall of each tool. Supplementary Figure S7 shows the impact of both the noise factor and the number of missed cleavages on the accuracy of pNovo 3, SMSNet, PointNovo and Casanovo. This demonstrates how a noise factor of at least 4 is already decreasing the prediction accuracy on spectra with no missing cleavage sites across all evaluated tools.

Furthermore, we investigated the relationship between peptide length, the number of missing cleavage sites and prediction accuracy (Figure 4). The prediction accuracy decreases from short peptides with few missing fragmentation sites to long peptides

with a high number of missing cleavages for each algorithm. The DL-based tools Casanovo, SMSNet and PointNovo show a higher prediction accuracy for peptides of a greater length compared with pNovo 3. PointNovo and SMSNet are able to learn sequence patterns of amino acids using their long short-term memory networks, which can overcome the issue of missing cleavage sites [18, 21]. The transformer-based approach allows Casanovo to process spectrum peaks as a whole and learn relationships between amino acids because of its self-attention mechanism. Spectrum graph-based methods show a lower peptide recall since missing fragmentation sites increase the complexity of possible peptide predictions [14]. McDonnell et al. [33] reported that the spectrum graph-based approach of Novor can correctly predict short subsequences of present fragment ions, in contrast to DeepNovo, which correctly predicts more complete peptides but fewer correct subsequences. However, even with all cleavage sites present, the evaluated algorithms only rarely identified correct peptides with a length of at least 18 amino acids. As expected, the number of correct predictions was higher for peptides below a size of 14 amino acids. Miscleavages lead to peptides of greater length, which would overall lower the prediction accuracy.

Following the influence of peptide length on the predictive performance, we compared the frequency of certain error types across multiple tools. We categorized incorrect peptide sequence predictions into 11 different error types and compared their relative amount between pNovo 3, SMSNet, PointNovo and Casanovo across all data sets (Table 3). We observed that for pNovo 3, SMSNet and PointNovo most errors were caused because of more than six wrongly assigned amino acids. Among the error types under 6 AAs, the inversion of the last 3 amino acids (PointNovo)



**Figure 2.** Total recall and precision of Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo across different enzymes on IgG1-Human-HC. (A) Recall at peptide level. (B) Recall at amino acid level. (C) Precision at amino acid level.

and the replacement of 1 AA by 1 or 2 AAs (SMSNet, pNovo 3) appear as the most frequent origins of incorrect peptide predictions. Conversely, Casanovo generated fewer predictions with a smaller number of mismatches in relation to pNovo 3, SMSNet and PointNovo. Again, this can be attributed to Casanovo's precursor filter, which results in a smaller fraction of errors, where more than 6 AAs were incorrectly predicted. The number of inversions was slightly lower on pNovo 3, demonstrating the advantage of a re-ranking framework for improved accuracy.

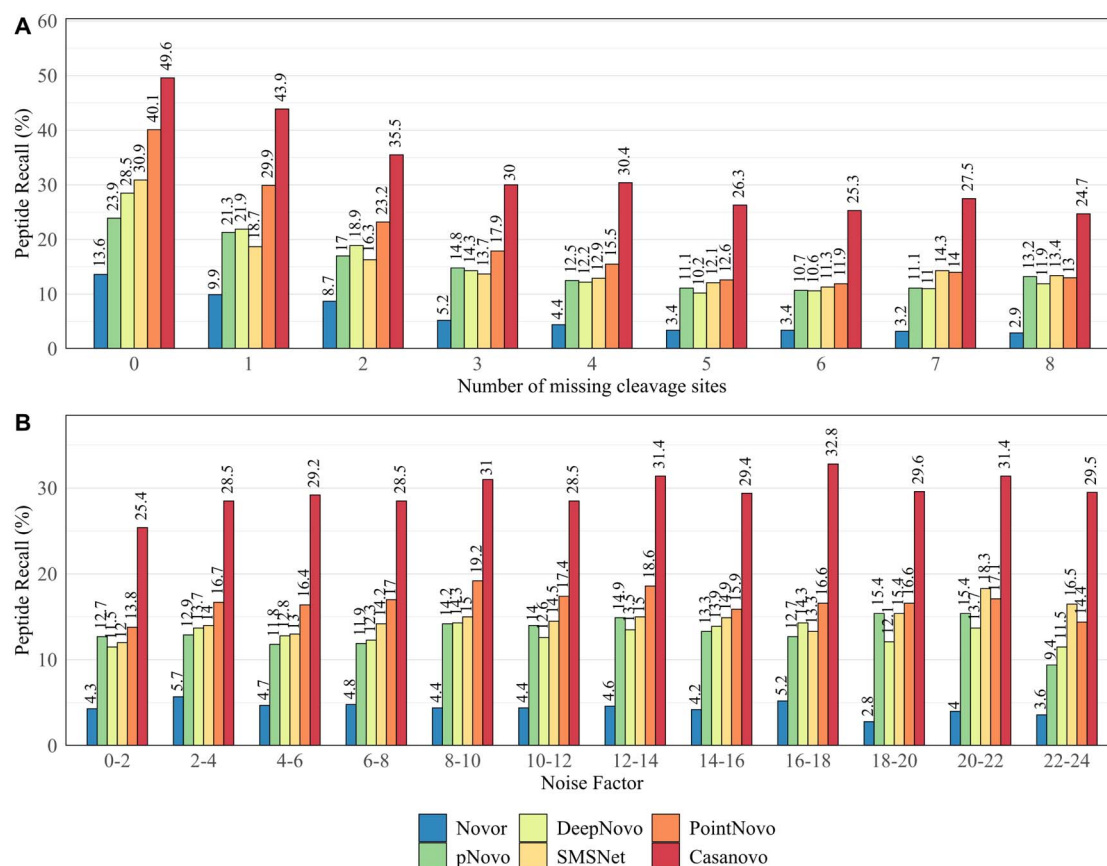
### Database-independent assembly of predicted peptide sequences

To validate the predictions of different peptide *de novo* sequencing tools on assembly level, we used the de Bruijn graph assembler ALPS, which generates several contiguous sequences (contigs) based on the *de novo* peptide results and their confidence scores [27]. We compared the longest constructed contig, the overall sequence coverage and the sequence accuracy for three antibody samples. As described in section Assembly of identified peptides, we considered only aligned contigs for the calculation of sequence accuracy. The light chains are 210–219 AAs long, whereas the heavy chains of our evaluated antibodies include over 440 AAs, which present a challenge for a complete sequence assembly. The longest constructed contig for the heavy chain of WlgG1 was generated by Casanovo, covering only 83 AAs (18.82%) of the protein sequence. On the light

chain of WlgG1, the results of Casanovo were concatenated to a contig, which covered 110 AAs (50.23%) of the entire sequence.

Since single contigs only cover a small region of the full-length protein, we evaluated the protein sequence coverage using a higher number of contigs for the light chain of IgG1-Human (Table 4). Here, we only evaluated SMSNet, PointNovo and Casanovo, since these three tools showed a higher AA recall and peptide recall across various enzymes and data sets compared with DeepNovo, Novor and pNovo 3. Combined with Casanovo, we were able to assemble 97.69% (IgG1) to 99.53% (Herceptin) of the whole antibody sequence with an accuracy of 94.47–95.26%. We observed a similar high sequence coverage for SMSNet (90.74–97.20%) and PointNovo (93.15–99.07%). Interestingly, we were able to achieve high coverage and accuracy on the light chain of WlgG1, although we only used the enzymatic data sets of chymotrypsin and asp-N.

Furthermore, we evaluated the assembly method on the heavy chains of our evaluated data sets (Supplementary Table S3). Here, we observed a lower sequence coverage and accuracy across all tools. Using Casanovo and ALPS, we achieved a sequence coverage of 76.39% (Herceptin) up to 93.72% (IgG1). We encountered multiple short overlapping contigs, which would make a full-length assembly more difficult without using additional tools. Moreover, these contigs include multiple mismatches, gaps, and were only partly aligned to the target sequence. Still, on the light chain



**Figure 3.** Total peptide recall of Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo across all data sets for different number of cleavage sites missing (A) and different noise factors (B) of the specific spectra.

**Table 3.** Error types made by *de novo* sequencing algorithms tools pNovo 3, SMSNet, PointNovo and Casanovo on the data sets of IgG1-Human, WlgG1-Mouse and Herceptin. Shown are the total number of predictions, total number of errors and the relative amount of 11 different error types for each algorithm. ‘Other’ includes errors that do not fall into any other categories, e.g. ‘2 AAs replaced by 4 AAs’

Type of sequencing error	pNovo 3	SMSNet	PointNovo	Casanovo
Number of total predictions	16 170	23 227	22 417	10 907
Number of total errors	14 860	20 554	19 240	5118
Inversion first 3 AAs (%)	5.5	5.3	5.0	11
Inversion last 3 AAs (%)	2.4	4.4	6.7	14
Inversion first and last 3 AAs (%)	0.2	0.4	1.0	0.8
1 AA replaced by 1 AA or 2 AAs (%)	10	5.4	5.6	27
2 AAs replaced by 2 AAs (%)	7.0	4.5	4.2	9.4
3 AAs replaced by 3 AAs (%)	6.1	3.7	3.6	6.1
4 AAs replaced by 4 AAs (%)	2.5	3.2	2.8	4.8
5 AAs replaced by 5 AAs (%)	2.6	2.6	2.1	2.3
6 AAs replaced by 6 AAs (%)	3.1	2.3	2.2	2.1
More than 6 AAs wrong (%)	44	57	55	10
Other (%)	14	10	10	9.9

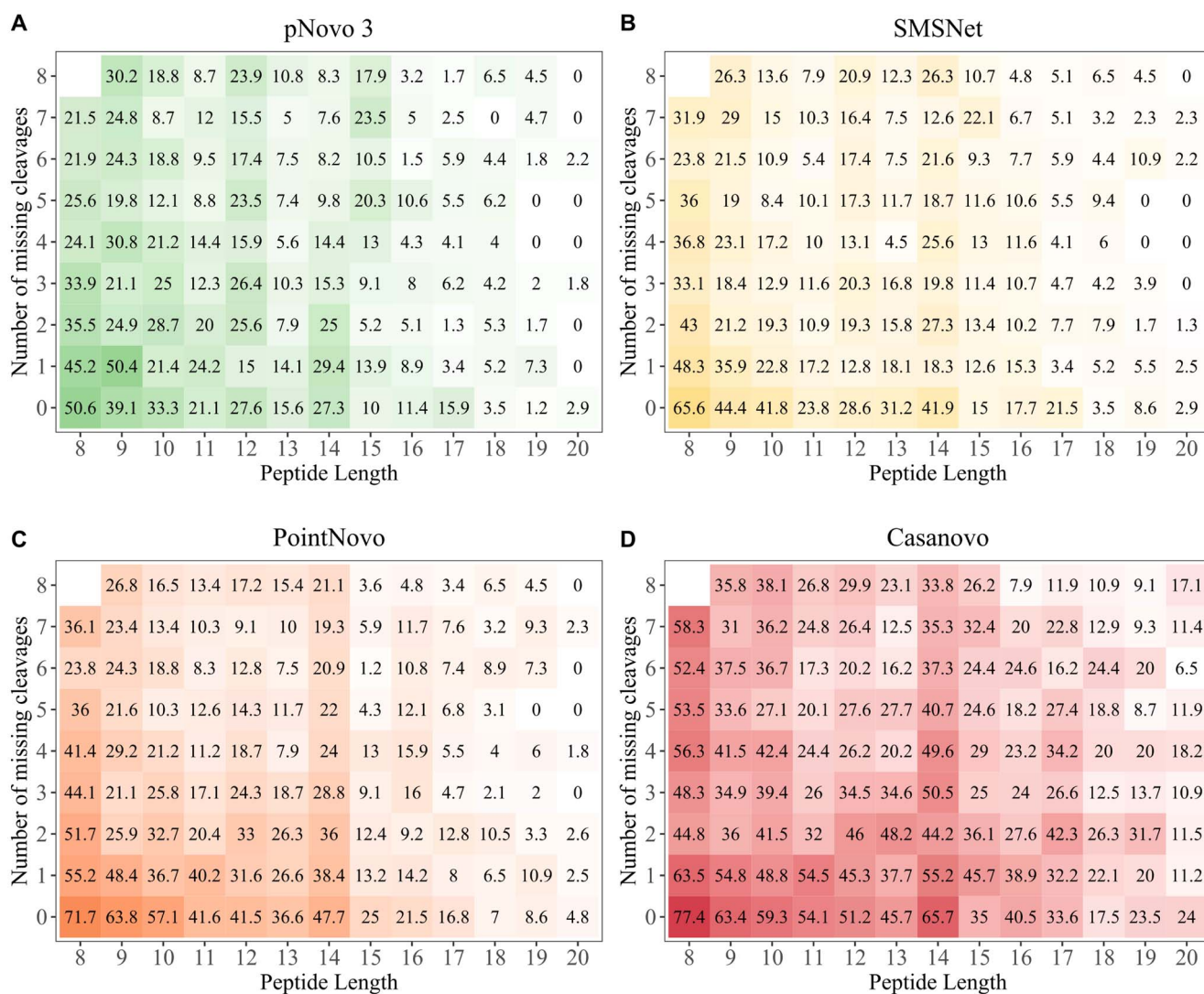
of Herceptin, we achieved a sequence coverage of 99.53% using Casanovo.

Despite the challenges of full *de novo* protein sequencing, we were able to correctly assemble functionally important subregions, namely the variable region and the CDRs, with the use of Casanovo and ALPS (Supplementary Table S5). We identified the corresponding CDRs for each antibody using the Natural Antibody database [55]. The CDRs were 100% correctly predicted on the light chain of IgG1. The heavy chain was correctly assembled except for a single misidentification on CDR3. The incorrect sequence

assignment included mismatches between amino acids with an identical mass (e.g. Q & GA; deamidated N & D; deamidated Q & E). The results on the variable regions and CDRs highlight the potential of DL-based *de novo* sequencing to identify unique antibody sequences.

## Discussion

In this study, we reviewed state-of-the-art *de novo* sequencing algorithms and applied them to the assembly of mAbs. We



**Figure 4.** Heatmap showing peptide recall for different number of missing cleavages (y-axis) and peptide lengths (x-axis). Higher peptide recall is shown in green for pNovo 3 (A), yellow for SMSNet (B), orange for PointNovo (C) and red for Casanovo (D). Lower peptide recall is displayed in white. Spectra are not distributed uniformly and the squares on the right and top of the plots include fewer spectra, since combinations of long peptides and a high number of missing cleavages (top right) occur less likely.

compared the performance of six recently developed and commonly used *de novo* peptide sequencing tools, namely, Novor, pNovo 3, DeepNovo, SMSNet, PointNovo and Casanovo.

Statistical analysis on amino acid and peptide levels revealed that the recently developed tools SMSNet, PointNovo and Casanovo achieved a high peptide recall on different enzymatic data sets (Figures 1 and 2; Supplementary Figures S2–S4). Similar to previous observations [21,56], DL-based algorithms predict a higher amount of correct peptide sequences compared with conventional spectrum-graph-based methods. A crucial factor for retrieving the correct peptide sequence is the resolution of the mass instrument and, consequently, the ability of *de novo* sequencing tools to make use of such high-resolution spectra [21]. For example, ambiguities between amino acids with similar mass (e.g. Q & K; oxidized M & F; AG & Q) cannot be resolved correctly on mass spectra with a wide fragment ion error tolerance of 0.1 Da and this makes MS/MS of higher resolution necessary [57]. As Qiao et al. pointed out, high-resolution spectra led to increased computational complexity for analyzing MS/MS data with *de novo* sequencing algorithms. DeepNovo and SMSNet

need to discretize spectra with a higher resolution parameter, which increases the computation and memory demand, whereas PointNovo and Casanovo can handle high-resolution spectra without increasing their computational complexity [21]. However, particular amino acids cannot be resolved even with spectra and tools with higher resolution (e.g. I & L; Q & AG; deamidated N & D). Here, additional methods are necessary to retrieve the correct amino acid sequence. Discrimination of the isomeric residues isoleucine and leucine cannot be achieved via MS/MS but require MS3 fragmentation [58,59].

Cross-enzyme performance is an important quality feature of *de novo* sequencing methods in bottom-up proteomics. Yet, most publications regarding *de novo* sequencing tools rarely address the predictive abilities of non-tryptic proteolytic enzymes and focus on tryptic data sets because of their wide availability and well-established usage [19,38]. Qiao et al. [21] observed that enzyme-specific models had a notable influence on the performance and recommended training a separate model for each enzyme. However, training different models for over six enzymes can be a demanding task, especially if the data are only partly available



**Table 4.** Summary of *de novo* assembly results on light chains of three antibody data sets using the *de novo* peptide sequencing tools SMSNet, PointNovo, Casanovo, and the de Bruijn assembler ALPS ( $k=7$ ). We used the Top 20 contigs to compare the length, coverage and accuracy of mapped contigs. Mapped contigs must be aligned to the reference protein sequence. The longest contig describes the maximum length of all generated contigs. Sequence coverage was calculated as the percentage of amino acids of the complete protein sequence that was covered by at least one contig. Accuracy was calculated as the percentage of all protein sequence calls that were labeled correctly

	IgG1 LC (216 AA)	WlgG1 LC (219 AA)	Herceptin LC (214 AA)
SMSNet			
Mapped contigs	10	5	8
Longest contig	51 (23.61%)	61 (27.86%)	67 (31.30%)
Sequence coverage	196 (90.74%)	200 (91.32%)	208 (97.20%)
Sequence accuracy	171 (87.24%)	190 (95.00%)	183 (87.98%)
PointNovo			
Mapped contigs	7	3	6
Longest contig	51 (23.61%)	108 (49.32%)	75 (35.05%)
Sequence coverage	205 (94.91%)	204 (93.15%)	212 (99.07%)
Sequence accuracy	187 (91.22%)	191 (93.63%)	190 (89.62%)
Casanovo			
Mapped contigs	7	4	4
Longest contig (AA)	65 (30.09%)	110 (50.23%)	105 (49.07%)
Sequence coverage (%)	211 (97.69%)	217 (99.09%)	213 (99.53%)
Sequence accuracy (%)	201 (95.26%)	205 (94.47%)	202 (94.84%)

or come from various sources with unequal experimental setups. Although our training data included mainly tryptic peptides, a relatively high number of non-tryptic peptides were identified by Casanovo, PointNovo, SMSNet and DeepNovo (Figure 2 and Supplementary Figure S3). Karunratanakul et al. [20] made a similar observation, where SMSNet was able to discover a large number of non-tryptic HLA-antigens, whereas 95% of its training data consisted of tryptic peptides. We conclude that DL tools can still be applied to different enzymatic data sets, although the performance will vary based on the cleavage pattern of the trained data set. In our opinion, the deployment of a higher number of enzyme-specific data sets and models would be beneficial for successfully applying *de novo* sequencing in proteomics. Furthermore, multienzyme DL models show the potential to improve the assembly of protein sequences [60].

Previous evaluations of *de novo* sequencing tools have observed an increased accuracy of these algorithms on simulated MS/MS spectra compared with real data sets [14, 33], suggesting that the bottleneck for *de novo* peptide identification lies in the quality of the provided data. As shown in our analysis, all evaluated tools show a higher peptide recall on spectra with fewer missing fragment ions (Figure 4). We observed that 90.51% of all spectra lacked at least one fragment ion. While newly developed tools demonstrate the potential of *de novo* sequencing, advanced post-processing steps are necessary to improve their accuracy. The DL-based tools SMSNet and PointNovo generated a higher number of completely incorrect peptides in comparison to pNovo 3 (Table 3). As Yang et al. [38] pointed out, DL models are directly learned from the MS/MS data and do not rely on well-designed features, which could help reduce the error frequency. Furthermore, the authors reported that even DL-based approaches have difficulties in distinguishing similar peptides with long-gapped subsequences, concluding that the quality of MS/MS data is a bottleneck of successful peptide prediction. However, the DL-based tool Casanovo displayed a lower number of incorrect predictions compared with pNovo 3, SMSNet and PointNovo, demonstrating the advantages of transformer-based models for predicting peptide sequences. The authors of Casanovo showed how a simple precursor mass filter yields much higher precision [22]. Moreover, it is worth

mentioning that several methods were published, discussing how to improve the encoder–decoder paradigm of DL tools in proteomics [56,61]. Fei pointed out that deep neural networks face difficulties on tandem mass spectra with incomplete fragment patterns. Multiple authors have confirmed that a considerable amount of *de novo* sequencing errors occurs at the N-terminal ends because of the absence and low intensity of fragment ions [62,63]. Hence, Fei developed a retrieve-and-revise framework to compensate for low-quality spectra. His peptide identification model, which relies on a reference database, was able to outperform current state-of-the-art algorithms [56]. Ge et al. proposed the use of deep residual shrinkage networks for their *de novo* sequencing method DePS to improve the accuracy on noisy spectra with missing fragmentation ions. Their implementation improved the extraction of features from MS/MS spectra and outperformed DeepNovoV2 on multiple data sets [64]. Liu et al. [65] used multiple temporal convolutional network blocks to improve the current state of *de novo* sequencing with their tools PepNet. Recently, the transformer-based approach DPST [66] showed an increased accuracy in predicting peptide sequences compared with DeepNovo, while reducing the model complexity and inference time.

Despite the ongoing effort and progress in *de novo* peptide sequencing, reliable protein assembly is still a demanding task. Our findings show that the ability of database-independent approaches of full-length protein assembly is limited even when using multiple contigs and different *de novo* sequencing tools (Table 4). The longest generated contig only covers at best 21.82% of the heavy chain and 50.23% of the light chain. Using Casanovo and ALPS, we accomplished a sequence coverage of 97.69–99.53% on the light chains of our evaluated antibodies. However, additional database tools or homology algorithms are necessary to correctly assemble multiple short contigs to complete antibody sequences. Homology search can be combined together with *de novo* sequencing to improve the discovery of protein sequence information and overcome problems caused by mass segment errors [26,67,68]. Commercial software packages such as PEAKS AB [27] and Supernovo [11] use antibody germline sequences as a starting point together with *de novo* sequencing results to identify

mAbs. Supernovo employs *de novo* peptide sequencing, database search, *in silico* genetic recombination and a final sequence assembly for an automatic antibody sequence prediction. Similarly, the publicly available software tool Stitch maps short peptide reads to user-defined templates for reconstructing monoclonal and polyclonal antibody sequences [69]. In addition to the before-mentioned homology tools, antibody-specific language models, such as AbLang, can help to restore missing residues of full protein sequences caused by sequencing errors without using a germline template sequence [70]. Thus, the development of publicly available frameworks and pipelines for automated assembly of *de novo* peptide sequencing results from recently developed algorithms would improve the reliable usability of *de novo* sequencing for full antibody assembly.

### Key Points

- A comprehensive review of *de novo* sequencing tools in proteomics is provided that aims to solve the challenge of antibody sequencing and subsequent assembly.
- Improved sensitivity of deep learning-based tools was found in comparison to classical *de novo* sequencing algorithms, such as spectrum graph-based algorithms, across various enzymatic data sets of antibodies.
- The number of missing fragmentation sites, noisy spectra and long peptide sequences poses a limit for all *de novo* sequencing tools.
- Database-independent assembly of light chains can be achieved up to a sequence coverage of 99.53% by using the de Bruijn assembler ALPS together with *de novo* peptide predictions from Casanovo.
- Further development of freely available and automatized pipelines for an accurate assembly of peptide predictions is necessary to successfully retrieve full antibody sequences.

### Author contributions

D.B. collected the data and performed the main analysis. T.M. and B.R. supervised the project. D.B. and T.M. wrote the manuscript with the help of B.R., G.T. and M.W. All authors reviewed and approved the manuscript.

### Data availability

Result files and Python code to reproduce the results in this study are available at Figshare ([doi.org/10.6084/m9.figshare.21394143](https://doi.org/10.6084/m9.figshare.21394143)).

### Acknowledgements

The authors would like to thank Natthan Ruengchaijatuporn (Chulalongkorn University) for his help with the retraining process of SMSNet.

### References

- Nelson PN, Reynolds GM, Waldron EE, et al. Demystified ...: monoclonal antibodies. *Mol Pathol* 2000;**53**(3):111–7.
- Goldman R. Antibodies: indispensable tools for biomedical research. *Trends Biochem Sci* 2000;**25**(12):593–5.
- Snapkov I, Chernigovskaya M, Sinitcyn P, et al. Progress and challenges in mass spectrometry-based analysis of antibody repertoires. *Trends Biotechnol* 2022;**40**(4):463–81.
- de Graaf SC, Hoek M, Tamara S, et al. A perspective toward mass spectrometry-based *de novo* sequencing of endogenous antibodies. *MAbs* 2022;**14**(1):2079449.
- Voskuil JLA. The challenges with the validation of research antibodies. *F1000Research* 2017;**17**(6):161.
- Bradbury A, Plückthun A. Reproducibility: standardize antibodies used in research. *Nature* 2015;**518**(7537):27–9.
- Weller MG. Quality issues of research antibodies. *Anal Chem Insights* 2016;**20**:11.
- de Marco A, Berrow N, Lebendiker M, et al. Quality control of protein reagents for the improvement of research data reproducibility. *Nat Commun* 2021;**12**(1):2795.
- Weller MG. Ten basic rules of antibody validation. *Anal Chem Insights* 2018;**1**(13):117739011875746.
- Sela-Culang I, Kunik V, Ofra Y. The structural basis of antibody-antigen recognition. *Front Immunol* 2013;**4**:302. <https://journal.frontiersin.org/article/10.3389/fimmu.2013.00302/abstract>.
- Sen KI, Tang WH, Nayak S, et al. Automated antibody *de novo* sequencing and its utility in biopharmaceutical discovery. *J Am Soc Mass Spectrom* 2017;**28**(5):803–10.
- Peng W, Pronker MF, Sniijder J. Mass spectrometry-based *de novo* sequencing of monoclonal antibodies using multiple proteases and a dual fragmentation scheme. *J Proteome Res* 2021;**20**(7):3559–66.
- Matallana-Surget S, Leroy B, Wattiez R. Shotgun proteomics: concept, key points and data mining. *Expert Rev Proteomics* 2010;**7**(1):5–7.
- Muth T, Renard BY. Evaluating *de novo* sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* 2018;**19**(5):954–70.
- Muth T, Hartkopf F, Vaudel M, et al. A potential golden age to come-current tools, recent use cases, and future avenues for *de novo* sequencing in proteomics. *Proteomics* 2018;**18**(18):1700150.
- O'Bryon I, Jenson SC, Merkley ED. Flying blind, or just flying under the radar? The underappreciated power of *de novo* methods of mass spectrometric peptide identification. *Protein Sci* 2020;**29**(9):1864–78.
- Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *arXiv* 2014. <https://arxiv.org/abs/1409.3215> 17 October 2022, date last accessed.
- Tran NH, Zhang X, Xin L, et al. *De novo* peptide sequencing by deep learning. *Proc Natl Acad Sci* 2017;**114**(31):8247–52.
- Tran NH, Qiao R, Xin L, et al. Deep learning enables *de novo* peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* 2019;**16**(1):63–6.
- Karunratanakul K, Tang HY, Speicher DW, et al. Uncovering thousands of new peptides with sequence-mask-search hybrid *de novo* peptide sequencing framework. *Mol Cell Proteomics* 2019;**18**(12):2478–91.
- Qiao R, Tran NH, Xin L, et al. Computationally instrument-resolution-independent *de novo* peptide sequencing for high-resolution devices. *Nat Mach Intell* 2021;**3**:420–5.
- Yilmaz M, Fondrie WE, Bittremieux W, et al. *De novo* mass spectrometry peptide sequencing with a transformer model. *Proceedings of the 39th International Conference on Machine Learning* 2022;**162**:25514–22.
- Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 2014;**5**(1):5277.

24. Rickert KW, Grinberg L, Woods RM, et al. Combining phage display with de novo protein sequencing for reverse engineering of monoclonal antibodies. *mAbs* 2016;**8**(3):501–12.
25. Savidor A, Barzilay R, Elinger D, et al. Database-independent protein sequencing (DiPS) enables full-length de novo protein and antibody sequence determination. *Mol Cell Proteomics* 2017;**16**(6):1151–61.
26. Guthals A, Clauser KR, Bandeira N. Shotgun protein sequencing with meta-contig assembly. *Mol Cell Proteomics* 2012;**11**(10):1084–96.
27. Tran NH, Rahman MZ, He L, et al. Complete de novo assembly of monoclonal antibody sequences. *Sci Rep* 2016;**6**(1):31730.
28. Mai ZB, Zhou ZH, He QY, et al. Highly robust de novo full-length protein sequencing. *Anal Chem* 2022;**94**(8):3467–75.
29. Ma B, Zhang K, Hendrie C, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;**17**(20):2337–42.
30. Zhang J, Xin L, Shan B, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 2012;**11**(4):M111.010587.
31. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags containing de novo sequencing error. *J Bioinform Comput Biol* 2005;**3**(03):697–716.
32. Pevtsov S, Fedulova I, Mirzaei H, et al. Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res* 2006;**5**(11):3018–28.
33. McDonnell K, Howley E, Abram F. The impact of noise and missing fragmentation cleavages on de novo peptide identification algorithms. *Comput Struct Biotechnol J* 2022;**1**(20):1402–12.
34. Cheng J, Wang L, Rive CM, et al. Complementary methods for de novo monoclonal antibody sequencing to achieve complete sequence coverage. *J Proteome Res* 2020;**19**(7):2700–7.
35. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2019;**47**(D1):D442–50.
36. Wang M, Wang J, Carver J, et al. Assembling the community-scale discoverable human proteome. *Cell Syst* 2018;**7**(4):412–421.e5.
37. Ma B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* 2015;**26**(11):1885–94.
38. Yang H, Chi H, Zeng WF, et al. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics* 2019;**35**(14):i183–90.
39. Zhou XX, Zeng WF, Chi H, et al. pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal Chem* 2017;**89**(23):12690–7.
40. Deutsch EW. File formats commonly used in mass spectrometry proteomics. *Mol Cell Proteomics* 2012;**11**(12):1612–21.
41. Adusumilli R, Mallick P. Data conversion with ProteoWizard msConvert. *Methods Mol Biol* 2017;**1550**:339–68.
42. Muth T, Weilnböck L, Rapp E, et al. DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J Proteome Res* 2014;**13**(2):1143–6.
43. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
44. Johnson RS, Searle BC, Nunn BL, et al. Assessing protein sequence database suitability using de novo sequencing. *Mol Cell Proteomics* 2020;**19**(1):198–208.
45. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;**20**(9):1466–7.
46. Barsnes H, Vaudel M. SearchGUI: a highly adaptable common interface for proteomics search and de novo engines. *J Proteome Res* 2018;**17**(7):2552–5.
47. Vaudel M, Burkhart JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 2015;**33**(1):22–4.
48. Levitsky LI, Klein JA, Ivanov MV, et al. Pyteomics 4.0: five years of development of a python proteomics framework. *J Proteome Res* 2019;**18**(2):709–14.
49. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res* 2010;**9**(3):1323–9.
50. Peng M, Taouatas N, Cappadona S, et al. Protease bias in absolute protein quantitation. *Nat Methods* 2012;**9**(6):524–5.
51. Low TY, vanHeesch S, Vandentoom H, et al. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep* 2013;**5**(5):1469–78.
52. Granholm V, Kim S, Navarro JCF, et al. Fast and accurate database searches with MS-GF+Percolator. *J Proteome Res* 2014;**13**(2):890–7.
53. Meyer JG, Kim S, Maltby DA, et al. Expanding proteome coverage with orthogonal-specificity  $\alpha$ -lytic proteases. *Mol Cell Proteomics* 2014;**13**(3):823–35.
54. Qi CR, Su H, Mo K, et al. PointNet: deep learning on point sets for 3D classification and segmentation. *arXiv* 2017. <https://arxiv.org/abs/1612.00593> 17 October 2022, date last accessed.
55. Krawczyk K, Buchanan A, Marcotili P. Data mining patented antibody sequences. *mAbs* 2021;**13**(1):1892366.
56. Fei Z. Retrieve and revise: improving peptide identification with similar mass spectra. In: David Leake, Indiana University (eds.) *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto, California USA: AAAI Press, 2021;**35**(17):14766–74. Online. <https://doi.org/10.1609/aaai.v35i17.17734> 26 November 2022, date last accessed. <https://aaai.org/Press/editorial.php>.
57. Medzihradsky KF, Chalkley RJ. Lessons in de novo peptide sequencing by tandem mass spectrometry: lessons in de novo peptide sequencing. *Mass Spectrom Rev* 2015;**34**(1):43–63.
58. Zhang W, Yang C, Liu J, et al. Accurate discrimination of leucine and isoleucine residues by combining continuous digestion with multiple MS3 spectra integration in protein sequence. *Talanta* 2022;**249**:123666.
59. Maibom-Thomsen S, Heissel S, Mørtz E, et al. Discrimination of isoleucine and leucine by dimethylation-assisted MS3. *Anal Chem* 2018;**90**(15):9055–9.
60. Gueto-Tettay C, Tang D, Happonen L, et al. Multienzyme deep learning models improve peptide de novo sequencing by mass spectrometry proteomics. *Bioinformatics* 2022. <https://biorxiv.org/lookup/doi/10.1101/2022.08.03.502594> 17 October 2022, date last accessed.
61. Fei Z. In: Derek Abbott (eds.) *Novel Peptide Sequencing With Deep Reinforcement Learning IEEE International Conference on Multimedia and Expo*. New York, NY 10016-5997 USA: IEEE, 2020; Online. <https://doi.org/10.1109/ICME46284.2020.9102827> 26 November 2022, date last accessed. <https://ieeaccess.ieee.org/editorial-leadership/senior-editors/>.
62. Blank-Landeshammer B, Kollipara L, Biß K, et al. Combining de novo peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics. *J Proteome Res* 2017;**16**(9):3209–18.
63. Miller SE, Rizzo AI, Waldbauer JR. Postnovo: postprocessing enables accurate and FDR-controlled de novo peptide sequencing. *J Proteome Res* 2018;**17**(11):3671–80.
64. Ge C, Lu Y, Qu J, et al. DePS: an improved deep learning model for de novo peptide sequencing. *arXiv* 2022;**16**:45–8. <https://arxiv.org/abs/2203.08820> 14 July 2022, date last accessed.

65. Liu K, Ye Y, Tang H. PepNet: a fully convolutional neural network for de novo peptide sequencing. In Review 2022. <https://www.researchsquare.com/article/rs-1341615/v1> 17 October 2022, date last accessed.
66. Yang Y, Hossain Z, Asif K, et al. DPST: de novo peptide sequencing with amino-acid-aware transformers. *arXiv* 2022. <https://arxiv.org/abs/2203.13132> 17 October 2022, date last accessed.
67. Liu X, Han Y, Yuen D, et al. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics* 2009;**25**(17):2174–80.
68. Pino L, Lin A, Bittremieux W. 2018 YPIC challenge: a case study in characterizing an unknown protein sample. *J Proteome Res* 2019;**18**(11):3936–43.
69. Schulte D, Peng W, Snijder J. Template-based assembly of proteomic short reads for de novo antibody sequencing and repertoire profiling. *Anal Chem* 2022;**94**(29):10391–9.
70. Olsen TH, Moal IH, Deane CM. *AbLang: an antibody language model for completing antibody sequences* *Bioinforma Adv* 2022;**2**(1):vbac046.