International Journal of Mass Spectrometry 471 (2022) 116723

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

A complete search of combinatorial peptide library greatly benefited from probabilistic incorporation of prior knowledge

Miroslav Hruska [a, b, *], Dusan Holub [a]

[a] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic
[b] Department of Computer Science, Faculty of Science, Palacky University, Olomouc, Czech Republic

ARTICLE INFO

Article history:
Received 14 April 2021
Received in revised form
29 July 2021
Accepted 6 October 2021
Available online 13 October 2021

Keywords:
Peptide detection
Tandem mass spectrometry
Computational analysis
Prior probability
Complete search
Theoretical modeling

ABSTRACT

The core of peptide detection in tandem mass spectrometry lies in associating fragment spectra with promising peptide candidates. We examined such detection in a synthetic combinatorial peptide library using four scoring metrics, against all theoretical peptides, and with a varying level of probabilistic prior knowledge—analyzing more than a trillion peptide-spectrum matches in total. Even after adjusting for peptide-length scoring bias, most MS/MS spectra had multiple at-least-as-good candidates as the correct peptide, showing that the highest spectral match was not a guarantee of correctness. As a remedy, we probabilistically integrated prior knowledge about expected cleavage behavior and expected peptide sequences into peptide scoring, reaching and even overcoming the performance of state-of-the-art *de novo* sequencing algorithms. Overall, we found that even partial and weak beliefs considerably improved peptide detection and are, in principle, generally applicable to any detection approach. Detection of peptides in a complete search thus often resulted in multiple admissible candidates near the maximal score, and the use of probabilistic prior knowledge substantially improved their discrimination.

1. Introduction

Shotgun proteomics is the leading technology for comprehensive explorative analyses of proteins in complex biological samples [1,2]. The core of peptide detection lies in the association of fragment spectra with peptides, and a variety of approaches for this purpose exist [3]. Traditionally, database search engines such as X! Tandem, MASCOT, or MS-GF+ calculate fragment masses of database peptides and match them against measured spectra [4–6]. Many research groups have recently focused on predicting fragment ion intensities instead of just fragment masses—an extension that improved peptide detection [7,8]. The best-performing spectra prediction models such as pDeep2 and Predfull utilize deep learning to learn complex molecular interactions during fragmentation and achieve high similarities of predicted and observed spectra [9,10]. Nonetheless, the database search has the drawback of considering a relatively small number of peptide candidates and a restricted set of modifications. For less restrictive analyses, open search enables reference-guided detection of peptides with any modification, including those with unknown masses [11,12]. Furthermore, its computational performance can be significantly improved using a fragment-ion index, as in MSFragger [13]. Recently, a hybrid tag-based approach in TagGraph enabled fast large-scale detection of peptides and their post-translational modifications [14]. If no prior knowledge about expected sequences is available or its use is not desirable, peptides can be detected using *de novo* sequencing [15,16]. From traditional machine learning approaches, Novor is one of the best-performing algorithms based on decision trees, with real-time sequencing performance [17]. Recent utilization of deep learning in DeepNovo and pNovo 3 further improved peptide sequencing *de novo* [18,19]. Nevertheless, the relatively low performance of *de novo* sequencing compared to the reference-guided searches makes them less useful in typical proteomics analyses.

Herein, we utilize a complete search strategy to obtain insights into several aspects of peptide detection, notably, the utility of prior probabilities for peptide detection. By prior probability of a peptide, we simply mean the probability that a randomly chosen peptide molecule from a sample of interest is the given peptide. Although the use of prior probabilities in ways similar to ours was investigated earlier [20–24], their direct incorporation into scoring while

* Corresponding author. Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic.
E-mail addresses: miroslav.hruska@upol.cz (M. Hruska), dusan.holub@upol.cz (D. Holub).

https://doi.org/10.1016/j.ijms.2021.116723
1387-3806/

considering various prior models is missing. Our study further shows that the use of prior probabilities also substantially limits the problem of an inflated search space [25,26], which affects database searches with many plausible candidates, such as in proteogenomics [27], or in immunopeptidomics [28,29]. Technically, our approach matched candidate peptides against a complete fragment-ion-indexed database [13], built separately for each precursor mass to allow fast calculation of spectral matches even for tens of millions of candidate peptides. The analysis enabled us to obtain the exact numbers of equal-or-better candidates, exact p-values, and general insights into the behavior of scoring in relationship to spectral characteristics. We compared the approach with state-of-the-art *de novo* sequencing algorithms and showed that even a simple scoring metric—the number of matching peaks—can perform reasonably well when integrated with cleavage-derived prior knowledge, and further improvements followed with a more involved scoring metric and more discriminatory prior models. Finally, we showed the ability to estimate the posterior probabilities of candidate peptides using Bayes' Theorem, allowing us to select peptides with the desired rate of false positives.

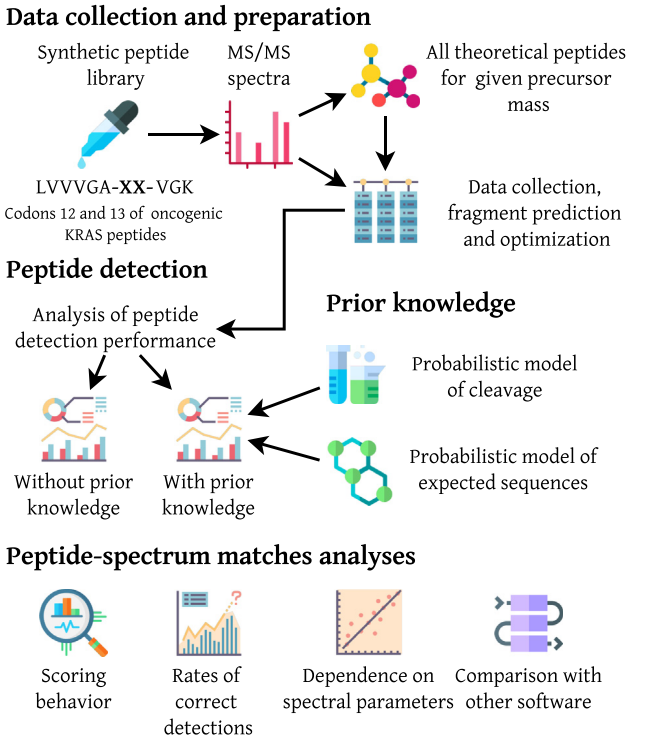


Fig. 1. Diagram of the overall analysis. 400 peptides from a synthetic peptide library of form LVVVGA-**XX**-VGK were each independently measured using LC-MS/MS. We selected 173 peptides from the library, totaling 3 078 spectra, such that there were at most 10^8 peptide candidates per spectrum and subjected them to a complete search. For each measured MS/MS spectrum, we constructed *in silico* all theoretical peptides for a given precursor mass and stored them in a database. We then predicted the theoretical fragments and built fragment-ion indexes to enable fast calculation of peptide-spectrum matches. Once the fragmentation data were prepared, we analyzed the scoring behavior of four scoring metrics. The analyses included: the exact numbers of equal-or-better candidates as the correct peptide, calculated p-values, the dependence of scoring on peptide length, rates of correct detections, and overall detection performance. Afterward, we incorporated varying levels of prior knowledge by modeling the expected cleavage behavior and by modeling the distance of candidate peptides to various expected reference sequences. We then calculated the posterior probabilities of individual peptides and investigated their behavior. For comparison, we analyzed the detection performance with commonly used search engines and *de novo* sequencing algorithms. In summary, we performed a complete search of MS/MS spectra of synthetic peptide library using multiple scoring metrics and evaluated their performance with and without probabilistic prior knowledge.

2. Results

2.1. Description of the synthetic library and the theoretical candidate peptides

Before we delve into the analysis of peptide-spectrum matches, let us first describe the synthetic peptide library and its relation to the theoretical peptides. The library is based on oncogenic KRAS peptides [30] and consists of 400 peptides of sequence pattern LVVVGA-**XX**-VGK (**XX** for any combination of amino acids, Fig. 1, Supplementary Table 1). As we aimed to analyze spectral matches for all theoretical peptides, we reduced the spectra in the library to those with a more manageable number of candidate peptides per spectrum (Fig. 2a). In particular, we picked MS/MS spectra with at most 10^8 candidates per spectrum, resulting in analyzing 3 078 MS/MS spectra from 173 peptides (Fig. 2a and b, Supplementary Table 2 and 3). The total number of candidate peptides for the selected part of the library was around two orders of magnitude less than for the whole library (5.42×10^9 vs. 3.44×10^{11}, Fig. 2c). Most of the analyzed MS/MS spectra originated from doubly charged precursors and were of medium intensity (double-charged: 2 144, triple-charged: 934; precursor intensities $Q_1/Q_2/Q_3$: 4.68×10^3/1.15×10^4/3.60×10^4, Fig. 2d). We focused on the spectra originating from doubly-charged precursor ions to simplify the analyses but included one analysis of triple-charged precursors for comparison. As the peptide library also contained some spectra unrelated to the peptides of interest, we considered only MS/MS spectra with a precursor mass within five parts-per-million (ppm) of the correct peptide to be the spectra corresponding to the peptide (Fig. 2e). The multiplicity of the spectra also allowed us to study the detection for varying precursor intensities, with 17.8 fragment spectra per peptide on average. In peptide-spectrum matching, we used the fragment tolerance of 0.3 on the *m/z* scale in accordance with the distribution of fragment mass differences (Fig. 2f). In summary, we selected a more computationally manageable part of the peptide library and analyzed it relative to all theoretical peptides and various prior probability models—an analysis that consisted of evaluating more than a trillion peptide-spectrum matches in total.

2.2. Most spectra had other at-least-as-good candidates as the correct peptide

The analysis of all theoretical peptides for a spectrum enabled us to obtain the exact numbers of at-least-as-good candidates as the correct peptide (Fig. 3a). Overall, we performed the analyses for four spectral match metrics computable using fragment-ion index [13], allowing for their fast calculation: the number of matching fragments (NMP); the rank-normalized sum of matching intensities (RPSM); the Hyperscore (HSPSM); and fragment frequency (FFPSM), a scoring metric we propose. NMP is a simple metric with certain desirable modeling properties, and RPSM is its extension that integrates fragment intensities into detection. Hyperscore is a metric originally introduced in X!Tandem [5] and its variation is in use in the open-search algorithm MSFragger [13]. FFPSM utilizes *a priori* distribution of expected fragments of all candidate peptides to suppress noise peaks, making it directly useable in a complete search (**Methods**). The numbers of candidates that matched the fragment ion spectra at least as well as the correct peptides were generally higher than one for each metric, indicating limited ability to detect uniquely the correct peptide (medians: 6 for NMP, 6 for RPSM, 4 for HSPSM, and 3 for FFPSM; double-charged precursors,

Supplementary Table 4). Note that rephrasing the previous in terms of p-values, we observed that the matches of the correct peptides were generally highly significant (medians of non-adjusted p-values: 3.95×10^{-7} for NMP, 3.67×10^{-7} for RPSM, 2.88×10^{-7} for HSPSM, and 2.34×10^{-7} for FFPSM). Even though the behavior was similar, FFPSM outperformed the remaining metrics in terms of equal-or-better matches (e.g., Wilcoxon $W = 234.00$, $p = 1.11 \times 10^{-12}$, $n = 173$ for HSPSM). Thus, although the analysis revealed other at-least-as-good candidates as the correct peptide for most spectra, scoring metrics exhibited significant differences in their behavior.

The number of at-least-good-candidates has a correspondence to E-values that are commonly reported by database search engines such as X!Tandem [4], Comet [31], or MS-GF+ [6]. E-value of a candidate peptide with score s refers to the expected number of peptides with score at-least-as-good as s, for a given size of the database used in the search. Note, however, that even very low E-Values do not guarantee the correctness of detection. To illustrate the point, we searched the peptide library against a database of peptides of pattern **VL**VVGA−−VGK (instead of **LV**VVGA−−VGK; no Isoleucines, 361 peptides). The peptide-spectrum matches had generally very low E-Values (median: 2.79×10^{-7} for MS-GF+), yet all of them were incorrect. This was because all candidate peptides were similar to the correct peptides—their matches were unlikely due to chance; however, that did not imply their correctness. Such a potentially misleading situation cannot happen in a complete search—there, the E-Value for a candidate peptide p equals the number of at-least-as-good candidates as p. Note that even if just one additional at-least-as-good candidate q exists, there is no preference to choose p over q unless there are other criteria or prior knowledge. Let us focus on the prior knowledge and consider two situations. First, suppose we know nothing about the sample; then, it is reasonable to treat all peptides as equally likely *a priori*. In such case, the other at-least-as-good candidate q is *a priori* as likely as p, and thus the posterior probability of p is at most one-half—often too low in practice. Now suppose we know that the sample is a trypsinized human sample. Then, the chance that some reference human tryptic peptide (e.g., p = **LV**VVGAGGVGK) is in the sample is many orders of magnitude higher than, say, the same peptide with two of its residues exchanged (e.g., q = **VL**VVGAGGVGK). Thus even if such other at-least-as-good candidate q exists, the posterior probability of p might be still high because q is unlikely *a priori*; if so, the posterior probability of q must be low. The relevance of existence of other at-least-as-good candidates thus depends on their prior probabilities. To summarize, the E-values in incomplete searches might be potentially misleading if there are non-searched peptides that are similar to the peptides in the search database, and which are not of sufficiently low prior probability.

2.3. B-ion ladders were more often matched at random compared to y-ion ladders

To provide an additional frame of reference to the scoring metrics analyzed, we also analyzed the detection by considering just theoretical b- and y-ion ladders (ʙNMP and ʏNMP, respectively). The detection based on either type of fragment ions was substantially worse compared to NMP, as indicated by the number of at-least-as-good candidates for the correct peptide (medians: 268 for ʙNMP, 28 for ʏNMP, and 6 for NMP). Note that matching against the b-ion ladders was less informative as the number of equal-or-better candidates was around one order of magnitude higher than for y-ion ladders (Wilcoxon $W = 574.00$, $p = 7.74 \times 10^{-25}$, $n = 173$). Similarly, the p-values for the correct peptides for ʙNMP were substantially higher (medians: 1.81×10^{-5} for ʙNMP vs. 1.70×10^{-6} for ʏNMP). In accordance, the b-ion ladders were more often matched at random compared to y-ion ladders (median of the average number of matching peaks per library peptide: 1.44 for b-ions vs. 0.80 for y-ions, Wilcoxon $W = 5076.00$, $p < 10^{-48}$, $n = 2\,144$). Therefore, a unit increase in the b-ion match was less relevant than a corresponding increase in the y-ion match.

Note that although the y-ions are generally more easily observed for tryptic peptides [32], the observed behavior resulted from a different phenomenon. In particular, the average number of matching b-ions for correct peptides was just slightly lower than the number of matching y-ions (7.76 vs. 7.84, resp.). On the other hand, the total number of matching peaks summed over all candidate peptides was almost twice as high over b-ion ladders if compared to y-ion ladders ($1.47\times/1.77\times/2.12\times$, $n = 2\,144$). Such b-to-y total-number-of-matching-peaks ratios also strongly correlated with the b-to-y ratios of at-least-as-good candidates over each spectrum (Spearman's $\rho = 0.88$, $p < 10^{-48}$, $n = 2\,144$; medians of ratios taken over each score level). The experimental fragments thus matched b-ion ladders more often and for more diverse candidate peptides, resulting in their lower ability to discriminate correct peptides.

2.4. A simple adjustment of peptide-length scoring bias increased the correct detections for the analyzed metrics by up to 13.2%

A close-up inspection of the tail of peptide matches distribution had often revealed peptides that were longer than the correct peptide (e.g., Fig. 5e). As longer peptides have more predicted fragments, scoring metrics based on matching fragments tend to give them an unfair advantage and such a bias impairs peptide detection [33]. Some search engines address the scoring bias (e.g., Comet [31] or MaxQuant [34]), while others do not, or at least not directly (e.g., X!Tandem [4] or MSFragger [13]). Still, when employing a target-decoy search strategy [35], the peptide length can be used as a feature to discriminate between target and decoy peptides (e.g., as in Percolator [36]), which performs a basic level of normalization. The target-decoy strategy is, however, not always applicable—particularly in a complete search because the target and decoy peptides for each spectrum are the same. We thus investigated the effect of length normalization on the detection of peptides in a complete search.

To examine this behavior, we first calculated correlations between the scores and peptide lengths (Fig. 3b). In general, the scores for each scoring metric had shown dependence on peptide length (medians of Spearman's ρ correlations: 0.138 for NMP, 0.124 for RPSM, 0.141 for HSPSM, and 0.183 for FFPSM). For instance, the average number of matching peaks for peptides of length nine was 1.75, compared to 2.10 for peptides with additional residue. Visually, we depicted average scores assigned by HSPSM, a metric similar to that in MSFragger [13], showing a clear yet mild growth of scores with peptide length (Fig. 3c). Afterward, we devised a simple adjustment, which subtracts the average score from the scores of candidate peptides of a given length (**Methods**). The adjustment suppressed the correlation of scores and peptide lengths (medians of correlations: 0.044 for NMP★, −0.027 for RPSM★, −0.041 for HSPSM★, and −0.032 for FFPSM★; Fig. 3b). Finally, we evaluated the detection performance of the length-adjusted metrics and found that they outperformed their non-adjusted counterparts. For instance, each length-adjusted metric outperformed the non-adjusted one in terms of equal-or-better candidates (e.g., 2/6/22 vs. 2/4/13 for NMP, Supplementary Table 5). As it often happened that the correct peptide had a maximal score, we focused on a more strict criterium—to identify the correct peptide *uniquely*, meaning that all incorrect peptides had a score strictly lower than the correct peptide. If the scoring

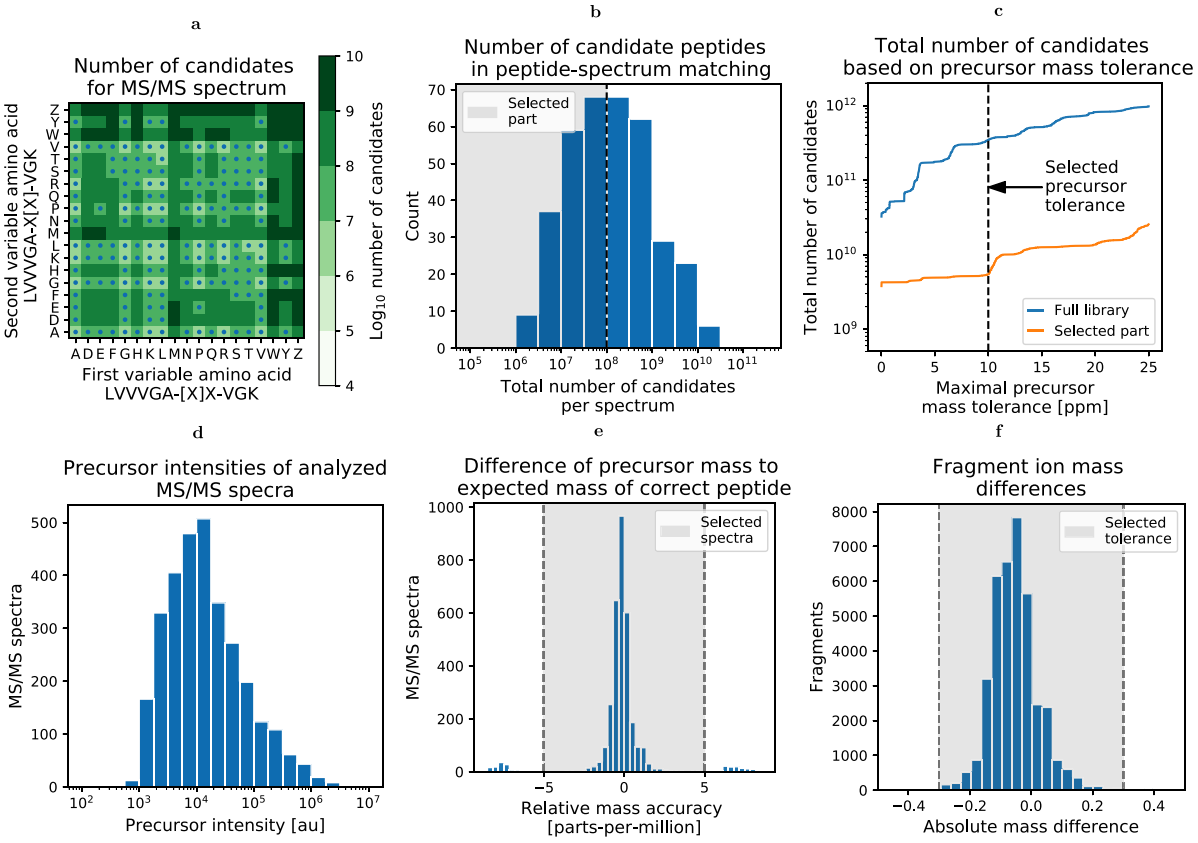


Fig. 2. Description of the synthetic peptide library. (a) The heatmap shows 173 peptides selected for the complete search analysis (blue dots). The peptides were selected such that the total number of peptide-spectrum matches per spectrum was at most 10^8 (see **b**). Note that we did not consider Isoleucines in the analysis; otherwise, the analyzed scoring metrics would never be able to identify uniquely and correctly any peptide. Further, we considered only carbamidomethylated cysteines (letter Z). **(b)** The distribution of the number of candidates for each peptide from the library. We selected only a subset of the library to simplify the analysis and suppress great demands on computational resources, mostly memory. Analyzing the full library would require around 100 TB of storage. Note that for each MS/MS spectrum of a peptide, we considered the same candidates (see **c**). **(c)** We considered all theoretical peptides within 10 parts-per-million (ppm) of the correct peptide's mass as theoretical candidates for scoring. Overall, the total number of candidates increased only mildly with the maximum allowed precursor tolerance. **(d)** Most spectra were of medium intensity, allowing us to study the detection in typical scenarios (median of precursor intensity: 1.15×10^4). **(e)** The distribution depicts the relative precursor mass difference of calculated and observed masses. We used each MS/MS spectrum from the peptide library within a precursor tolerance of five ppm of the correct peptide as an MS/MS spectrum of the correct peptide. **(f)** The plot depicts the distribution of fragment mass differences. We matched the predicted fragment ions to the closest fragment in MS/MS spectrum and showed such distribution within 0.5 on *m/z* scale, selecting 0.3 as a fragment mass tolerance.

metric detected the peptide correctly and uniquely, we referred to such a situation as a *correct interpretation of a spectrum*. The total increase in terms of correctly interpreted spectra reached 13.2% for NMP*, 9.3% for RPSM*, 10.7% for HSPSM*, and 3.2% for FFPSM*. The length-based adjustment of scoring metrics thus moderately improved peptide detection performance.

2.5. A spectral match approach rarely interpreted spectra correctly in a complete search

To examine the ability to detect the correct peptide sequence, we analyzed the detection performance based on spectral characteristics. First, we started with the analysis of 2 144 spectra of doubly charged precursor ions. Overall, the rates of correctly interpreted spectra were rather low for each scoring metric (Fig. 3d). Considering the worst and the best scoring metrics, NMP interpreted 13.4% of spectra correctly, while the FFPSM* 21.1% of spectra (57.3% more). The ability to correctly interpret spectra increased with the precursor intensity for all scoring metrics (e.g., Spearman's $\rho = 0.30$, $p = 7.78 \times 10^{-46}$, $n = 2\ 144$ for NMP). For instance, restricting the analysis to precursors of intensity at least 5×10^4 resulted in 30.2% correctly interpreted spectra for NMP, and 49.6% for FFPSM*. The scoring metrics thus had only limited ability to interpret the spectra correctly but this ability increased with the intensity of precursors.

For comparison, we have included complete analyses of double-charged precursor spectra performed using three popular search engines: MSFragger, Comet, and MS-GF+ (**Methods**). MSFragger resulted in 17.2% correct detections, similar to the metric HSPSM that aimed to mimic its behavior (17.4%). Note that the length-adjusted HSPSM* improved the detection to 19.3%, and the adjustment would most likely benefit MSFragger in a similar way. Comet reached 21.3%, similar to that of FFPSM* (21.1%). Finally, the highest performance was achieved by MS-GF+, reaching 23.9% of correctly interpreted spectra, and thus outperforming other scoring metrics and search engines in this respect.

Finally, we analyzed 934 triple-charged precursor spectra from 66 peptides. Overall, if multiply-charged ions were not allowed in matching, there were no spectra interpreted correctly by any of the scoring metrics. Similarly, neither Comet nor MSFragger was able to interpret any spectrum correctly (MS-GF+ did not allow restricting the maximal fragment charge). If we allowed multiply-charged fragment ions, the performance improved but very mildly (0.4% of correct and unique detections for NMP, 0.2% for MS-GF+, 0.3% for Comet, and 0.2% for MSFragger). The complete search of triple-charged spectra had thus very rarely interpreted the spectrum

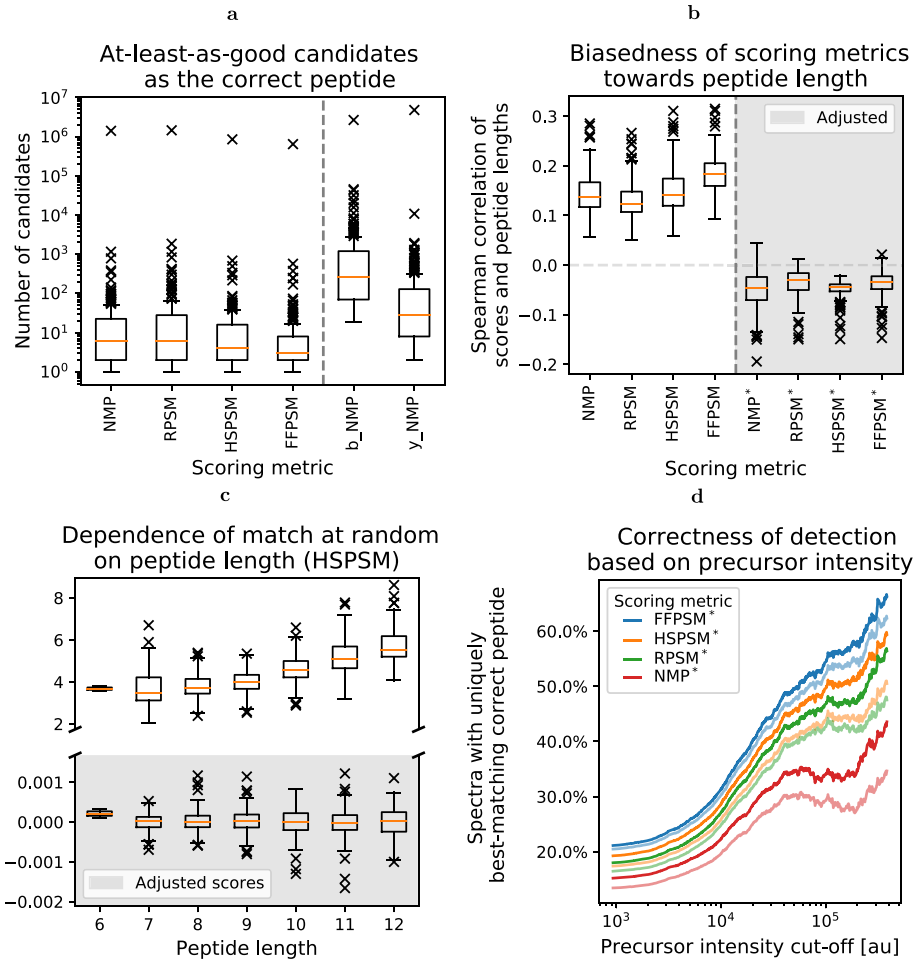


Fig. 3. Behavior and performance of scoring metrics. **(a)** The numbers of candidates that were score-wise at least as good as the correct peptide were generally higher than one for each scoring metric. The boxplot depicts the medians of the number of such candidates per peptide from the library (over all its spectra). **(b)** The scores of candidate peptides correlated with their lengths, thus showing a bias towards longer peptides. Substracting the average score per length suppressed the length effect (right side, "Adjusted"). The boxplot shows the averages of Spearman's ρ correlations per peptide from the library (over all its spectra). **(c)** Depiction of average scores calculated using HSPSM and HSPSM*. Although the average scores increased using HSPSM, they were close to zero in HSPSM*. The boxplot shows the average scores per peptide from the library (over all its spectra). **(d)** The ability to detect the correct peptide uniquely increased with the intensity of precursor ions. For instance, for the 10% of spectra with the most intense precursors, the best-performing FFPSM* interpreted correctly and uniquely almost double the spectra compared to NMP (55.9% vs. 28.5%, 1.96×). The lines with lower opacity show the behavior of metrics without the length adjustment.

correctly for the analyzed dataset, providing room for further investigations.

2.6. Probabilistic incorporation of expected cleavage behavior improved complete search performance

To investigate the effect of expected cleavage behavior on detection performance, we integrated spectral matches with cleavage-derived prior probabilities of peptides. Foremost, note that the majority of theoretical candidates did not conform to the cleavage of trypsin at C-terminal (median: 83.5%, Fig. 4a). Similarly, most of the theoretical candidates had at least one missed cleavage (overall: 91.4%, Fig. 4b). Naturally, one would expect that incorporation of these characteristics into peptide detection will improve its performance. Note that the number of missed cleavages and conformance to C-term specificity can also be integrated into peptide detection when employing target-decoy strategy [35]; this, however, remains inapplicable for complete search (as in the case of peptide-length normalization).

We thus integrated the expected cleavage behavior by assigning different prior probabilities to peptides based on their conformance to cleavage. In practice, we integrated the NMP and FFPSM* metrics with prior probabilities of peptides and derived the posterior probabilities using Bayes' Theorem (Fig. 4c and d; Supplementary Tables 7, 11–13; **Methods**). Note that even reformulation of the metrics into their probabilistic form substantially increased the areas under the precision-recall curves even for the uniform prior (AUC: 0.162 → 0.229 for NMP$_\mathrm{p}$, 0.251 → 0.305 for FFPSM$^\star_\mathrm{p}$). The result shows that filtering using probabilities instead of the raw scores generally allowed selecting more candidates at a given false-positive rate. The use of prior distribution based on the number of missed cleavages further improved the detection (AUC: 0.229 → 0.265 for NMP$_\mathrm{p}$, 0.305 → 0.348 for FFPSM$^\star_\mathrm{p}$; correctly interpreted spectra: 13.4% → 16.5% for NMP, 21.1% → 25.3% for FFPSM$^\star_\mathrm{p}$; multiplication of prior probability by 0.1 with each missed cleavage). Similarly, multiplying the prior probabilities by 0.001 for peptides with non-specific cleavage at C-terminal again improved the performance to an even greater degree (AUC: 0.229 → 0.289 for NMP$_\mathrm{p}$, 0.305 → 0.375 for FFPSM$^\star_\mathrm{p}$; correctly interpreted spectra: 13.4% → 17.5% for NMP, 21.1% → 26.3% for FFPSM$^\star_\mathrm{p}$). Reformulation of scoring into probabilistic settings and

incorporation of expected cleavage behavior thus substantially improved peptide detection performance.

As we analyzed the behavior of all theoretical peptides, we also compared the performance of spectral match metrics to *de novo* sequencing algorithms (Fig. 4e, Supplementary Table 8). The highest performance was reached by Novor (AUC: 0.349), followed by DeepNovo (AUC: 0.298), and finally PepNovo (AUC: 0.281). It was somewhat surprising that the use of the probabilistic version of simple FFPSM* with C-term cleavage model outperformed these algorithms (AUC: 0.375). Note that the use of the C-term modeling was justified in the comparison because all these algorithms were run with trypsin as a protease, thereby providing advantages to tryptic peptides. On the other hand, these algorithms could also be directly trained when appropriate data are available and are thus likely to reach higher performance in such circumstances. That being said, the simple scoring metrics are almost without parameters, so their potentially high performance provides room for future investigations as they are likely to be of similar performance across datasets. Detection using FFPSM* thus reached better performance than the complex peptide-scoring *de novo* sequencing algorithms on this synthetic combinatorial peptide library dataset.

2.7. Better-matching incorrect peptides had a high editing distance to the correct peptide

To get a closer look into the behavior of scoring metrics, we depicted spectral match distributions for a particular MS/MS spectrum (Fig. 5a–d, Supplementary Table 9). Although the details of each distribution differed, all of them had shown similar trends. In this example, each scoring metric assigned a near-maximal score to the correct peptide (17/18 for NMP, 1464/1523 for RPSM, 30.70/33.03 for HSPSM, and 142.50/149.38 for FFPSM). In accordance, a few strictly better matching candidates were visible at the right tail of the distribution (4 for NMP, 2 for RPSM, 8 for HSPSM, and 4 for FFPSM). The spectral matches of the correct peptide were thus very high for all metrics; however, a couple of candidates still matched the spectrum better.

Even though the scoring metrics were generally inadept in detecting the correct sequence uniquely, other plausible candidates were sequence-wise far from the correct sequence. The Fig. 5e shows candidate sequences, scores, and editing distances to the correct peptide (ΔSeq) for the analyzed spectrum. The better matching candidates were of the following median ΔSeq: 3.5 for NMP, 3.0 for RPSM, 4.0 for HSPSM, and 3.5 for FFPSM. The table thus illustrates that even though some candidates had a slightly better spectral match, their sequences might be unlikely, e.g., relative to

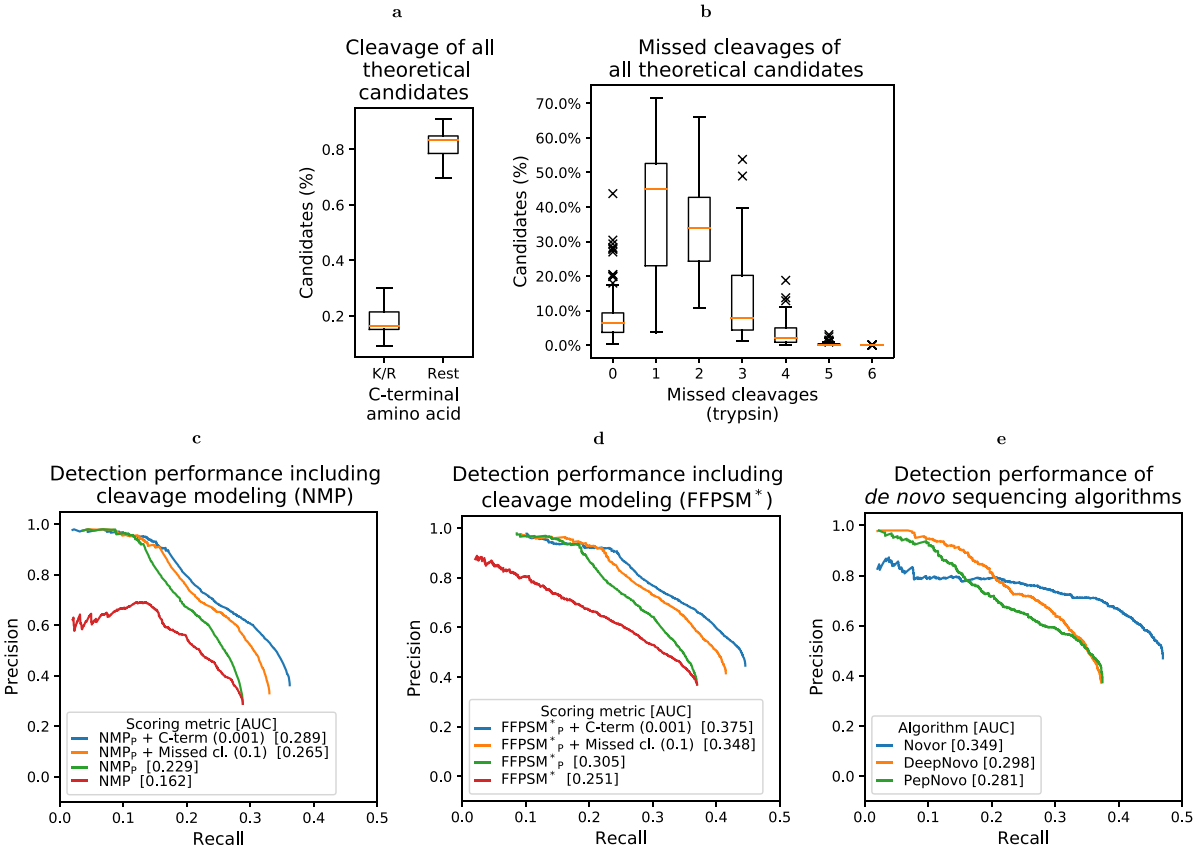
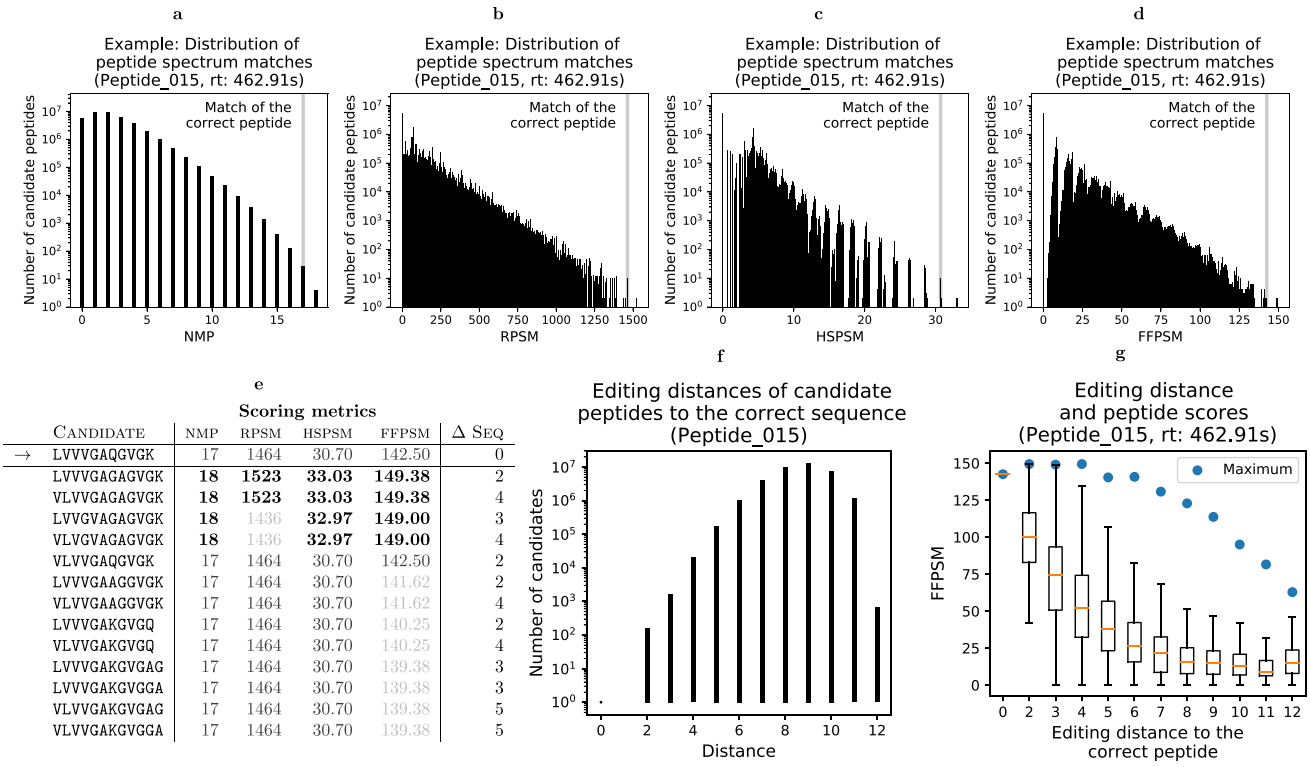


Fig. 4. Probabilistic integration of cleavage into peptide detection. **(a)** The proportion of theoretical candidates with C-terminal lysine (K) or arginine (R) represented around one-sixth of all theoretical candidates (median: 16.5%). **(b)** Most of the theoretical candidates had at least one missed cleavage (overall proportion: 91.4%). **(c)** Reformulating the NMP metric into its probabilistic version NMP_p improved the detection performance by allowing to select peptides at much higher precision. The incorporation of missed cleavages and C-term specificity further improved detection performance. The numbers in parentheses signify the decrease in prior probabilities of peptides (0.001 in C-term for non-specific cleavage and multiplication by 0.1 for each missed cleavage). **(d)** Similarly as in **c**, the probabilistic version of FFPSM* outperformed its non-probabilistic counterpart. Further improvements followed with the probabilistic modeling of cleavage behavior. **(e)** The performance of probabilistic versions of scoring metrics was on par with *de novo* sequencing algorithms, especially when using the modeling of cleavage (see **b** and **c**).



	Candidate	NMP	RPSM	HSPSM	FFPSM	Δ Seq
→	LVVVGAQGVGK	17	1464	30.70	142.50	0
	LVVVGAGAGVGK	**18**	**1523**	**33.03**	**149.38**	2
	VLVVGAGAGVGK	**18**	**1523**	**33.03**	**149.38**	4
	LVVGVAGAGVGK	**18**	1436	**32.97**	**149.00**	3
	VLVGVAGAGVGK	**18**	1436	**32.97**	**149.00**	4
	VLVVGAQGVGK	17	1464	30.70	142.50	2
	LVVVGAAGGVGK	17	1464	30.70	141.62	2
	VLVVGAAGGVGK	17	1464	30.70	141.62	4
	LVVVGAKGVGQ	17	1464	30.70	140.25	2
	VLVVGAKGVGQ	17	1464	30.70	140.25	4
	LVVVGAKGVGAG	17	1464	30.70	139.38	3
	LVVVGAKGVGGA	17	1464	30.70	139.38	3
	VLVVGAKGVGAG	17	1464	30.70	139.38	5
	VLVVGAKGVGGA	17	1464	30.70	139.38	5

Fig. 5. Distribution of peptide-spectrum matches and editing distances to the correct sequence. (a–d) The distributions show peptide-spectrum matches for all theoretical candidates for a particular spectrum. The spectrum was randomly selected from spectra that have a near-maximal score in each metric. **(e)** The table details the peptide-spectrum matches near the maximal score. The ΔSeq shows the editing distance of the candidate sequence to the correct sequence LVVVGAQGVGK (see also **f**). **(f)** The distribution of editing distances to the correct sequence LVVVGAQGVGK for all candidate peptides. Note that no peptide has ΔSeq of 1—if a sequence was edited by one, it would not fit into the tight precursor mass tolerance window. Nevertheless, that would change if we also considered Isoleucines in our analysis (e.g., the mass of a peptide after editing L → I remains the same). **(g)** The scores of peptides increased as the sequence was getting closer to the correct peptide. However, the scores were still generally far from the score of the correct sequence.

expected reference sequences. In general, only a handful of candidates were sequence-wise close to the correct peptide (Fig. 5f). In this example, 99.9951% of candidates were of editing distance larger than 3, far from the correct sequence. On the other hand, peptides close to the correct sequence were of substantially higher scores yet generally had much lower scores than the correct peptide (Fig. 5g). For instance, peptides of distance 2 had a score of 95.53 on average, far from the score of the correct sequence (142.50, FFPSM). In summary, the results thus suggested that modeling the candidates' distance, if appropriate at given circumstances, should significantly improve peptide detection.

2.8. Modeling of distance to the expected reference sequence dramatically improved peptide detection performance

To examine the relevance of *a priori* knowledge of correct peptides' sequence pattern, we analyzed the detection performance for various distance-based prior models. We modeled the prior probabilities using a *distance factor* (DF)—a number in the $(0, 1\rangle$ interval that signifies the multiplicative decrease in prior probability with a unit increase in editing distance to the reference sequence. Note that the posterior probabilities assigned to the correct peptides *without* utilizing prior knowledge were generally low (0.01/0.11/0.43, double-charged precursors, NMP_p, Fig. 6a, Supplementary Table 10). The calculated probabilities indicated that the unaided detection of peptides in complete searches using NMP_p metric would not be sensitive enough in practice. However, using the distance factor of just 0.1 substantially raised the probabilities to 0.42/0.86/0.98 (reference sequence: LVVVGA––VGK). In accordance, the rates of correctly interpreted spectra increased from 13.5% to 72.2%. Lower distance factors further increased the posterior probabilities (i.e., 0.76/0.97/1.00 for DF = 0.01, and 0.92/0.99/1.00 for DF = 0.001). Therefore, putting even mild importance on the distance of peptides to the reference sequence significantly raised the posterior probabilities for correct peptides.

To evaluate the overall detection performance, we compared the total number of interpreted spectra at a particular false positive rate (Fig. 6b). Overall, the use of even a small degree of prior knowledge significantly improved detection. For instance, considering precision of 90.0% resulted in 13.5% recall without prior knowledge, but increased to 69.2% for DF = 0.1. Similarly, the areas under the precision-recall curves were much higher when utilizing prior knowledge (e.g., AUC = 0.741 for DF = 0.1, compared to AUC = 0.229 without using prior knowledge, NMP_p). In summary, the detection was largely improved even with high distance factors if the prior probabilities were based on the distance to the expected reference sequence.

Afterward, we analyzed detection behavior for variably incomplete reference sequences (Fig. 6c; Supplementary Table 10–14). The highest posterior probabilities were assigned to the correct peptides when the expected reference sequence corresponded directly to the pattern of the peptides in the library (Pr = 0.42/0.86/0.98, LVVVGA––VGK, DF = 0.1). The use of the actual human reference sequence LVVVGAGGVGK resulted in moderately decreased probabilities, although it corresponded better to a real-world situation (Pr = 0.15/0.73/0.97, DF = 0.1). Overall, the posterior probabilities decreased with the number of missing amino acids around the variable part of the sequence (medians: 0.50, 0.45,

and 0.23 for 4, 6, and 8 missing amino acids, resp.; DF = 0.1). Finally, the posterior probabilities were the lowest when nothing except the length of the sequence was expected (median: 0.15, DF = 0.1). In summary, even mediocre completeness of the expected reference sequence substantially improved peptide detection compared to no prior knowledge.

2.9. Posterior probabilities were close to their desired behavior

Next, we analyzed the behavior of posterior probabilities to see if they follow their intended behavior. First, we chose the best candidates for each spectrum and compared the sum of their probabilities to the number of correct detections for various distance factors (Fig. 6d, Supplementary Table 7). Although the behavior for DF $\leq$ 0.1 followed the desired behavior closely, it had shown a particular phenomenon: an underestimation of probabilities for high distance factors (i.e., 0.5 and 0.9). The reason is that when multiple peptides have the same peptide-spectrum match, ratios of their posterior probabilities are equal to the ratios of their prior probabilities in our model. In these circumstances, the correspondence between the prior model and the data becomes important to obtain accurate posterior probabilities. For illustration, suppose there are two candidate peptides p_a and p_b, both of a maximal score. Let p_a be of distance 2 to the pattern LVVVGA−−VGK, p_b be of distance 3, and suppose that p_a is the correct peptide. At DF = 0.9, the posterior probability of p_b will be 0.9× that of p_a, thus just slightly lower than of p_a. For further simplicity, suppose p_a and p_b are the only candidates for the analyzed spectrum. As a result, p_a will have a posterior probability slightly above 0.5, and p_b slightly below 0.5 (i.e., 0.526 vs. 0.474, respectively). Because we are interested in the *best* candidates per spectrum, we will always select the one with the slightly higher probability. As all correct peptides are of distance 2 to the pattern LVVVGA−−VGK, peptides of the same match but of higher distance will have just slightly lower posterior probabilities than the correct peptides. We refer to such situations as *prior-induced underestimation* as they result from the lack of sufficient correspondence between the prior and the analyzed library at high distance factors. On the other hand, such situations do not exist at DF = 1, because these small numerical differences disappear—both peptides will have a probability of 0.5, and there would be thus no numerical advantage for the selection of the *best* candidate. The

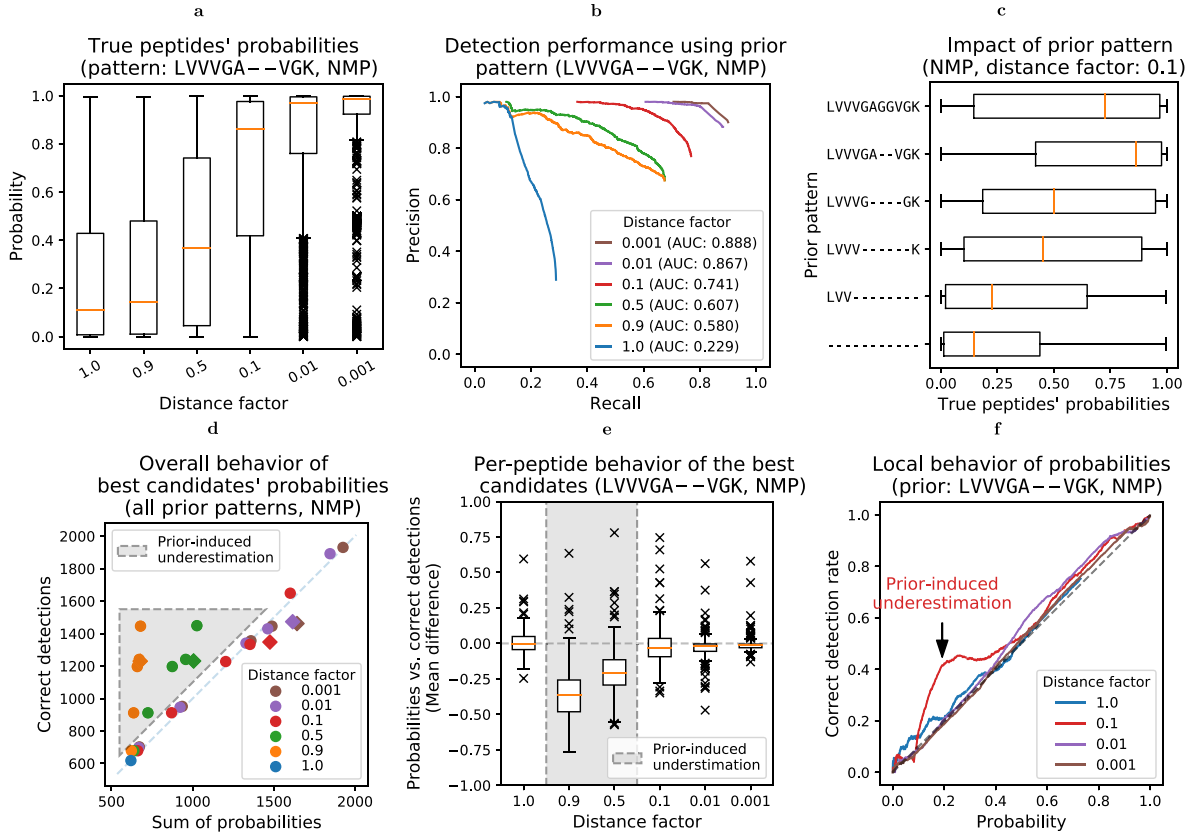


Fig. 6. Integration of spectral matches with expected peptide sequences. **(a)** Lack of prior knowledge resulted in low posterior probabilities assigned to the correct peptides (median of Pr = 0.11). Note that no prior knowledge corresponds to the distance factor (DF) of 1—no penalization with an increase in distance to the prior pattern (ΔSeq). On the other hand, using low DF = 0.001 assigned very high probabilities to the correct peptides (median of Pr = 0.986). **(b)** The use of prior knowledge of limited importance substantially raised the detection performance (e.g., from AUC = 0.229 for no prior knowledge to AUC = 0.741 for DF = 0.1). Strengthening the relevance of prior knowledge further improved detection performance. **(c)** The use of partially incomplete expected reference sequences still improved the peptide detection. For instance, the expected peptide sequence pattern of LVVV−−−−−−K resulted in mediocre probabilities at a mild distance factor (Pr = 0.10/0.45/0.89, DF = 0.1). **(d)** The expected number of correct detections corresponded well with the total number of correctly detected peptides (individual points in the plot represent different partial structures as seen in **c**). The prior-induced underestimation is a result of a situation when multiple candidate peptides have the same match, putting more relevance on prior probabilities. These, in turn, did not correspond well to the peptide library at high distance factors (DF > 0.1). Note that the behavior for LVVVGA**GG**VGK prior resulted in a slight overestimation of the probabilities at sufficiently low distance factors (diamonds). The reason is that the prior does not properly correspond to the peptide library structure and thus eventually starts forcing the incorrect behavior (at low enough distance factors). **(e)** The probabilities of the best candidates analyzed for each of the 173 peptides separately have shown a consistent behavior (see also **d** for prior-induced underestimation). **(f)** The posterior probabilities of all candidates were generally close to their desired behavior, indicating their correct estimation across a wide probabilistic range (see also **d** for prior-induced underestimation).

underestimation of probabilities thus resulted from an elevated dependence on prior model, which did not correspond well enough to the peptide library at high distance factors.

With the previous in mind and restricting the analysis to DF $\leq$ 0.1, the sums of probabilities were close to the actual numbers of correct detections ($y = 0.984\times$, $R^2 = 99.76\%$, ordinary least squares regression). Afterward, we investigated a more localized behavior of the probabilities. First, we studied the behavior for each of the library peptides separately to see if there are systematic deviations from the desired behavior (Fig. 6e, LVVVGA−−VGK, Supplementary Table 14). Except for the prior-induced underestimation for high distance factors, the differences in the average probabilities of the best candidates and the average rates of correct detections were close to zero (e.g., medians: 0.007 for the uniform prior, and −0.029 for DF = 0.1). The results thus suggested that the calculated probabilities were also reliable on a more individual basis. Finally, we examined the average behavior of probabilities for all candidates (Fig. 6f, Supplementary Table 14). Except for the prior-induced underestimation, the posterior probabilities were close to the average rates of correct detections, indicating their accurate estimation over the whole probabilistic range. Nevertheless, we observed a prior-induced underestimation resulting from a more complex scenario at DF = 0.1, showing the relative relevance of having a proper correspondence between the prior probabilities and the analyzed peptides. The posterior probabilities thus showed appropriate behavior for a uniform prior and for low distance factors, with potential complications caused by equal peptide-spectrum matches and corresponding elevated dependence on a correct model of prior probabilities. In summary, a reasonable behavior of posterior probabilities for a multitude of settings indicated the possibility to select peptides at a particular rate of false positives in a complete search integrated with prior probabilities.

3. Discussion

Our study examined the complete search strategy and the importance of scoring metrics and probabilistic prior knowledge for peptide detection. Although the complete search is feasible only for precursors of low-to-medium mass, our analysis offered various insights into the peptide scoring, such as the dependence of scoring on peptide length (Fig. 3b and c), and showed a clear utility of peptide prior probabilities in peptide detection (Fig. 4c and d). From a theoretical perspective, a complete search ensures that no candidate is missed and thus allows normalizing probabilities of candidates to sum to one. As many viable candidates usually reside around the maximal score, the prior knowledge can help discriminate between them—the highest match is typically not a sufficient guarantee of correctness (Fig. 5e). Furthermore, even though the search space of candidates was quite large in our analyses—maximal with 9.88×10^7 candidates—the use of discriminative prior probabilities anyway resulted in high posterior probabilities for correct peptides (Fig. 6a). The use of prior probabilities thus helped leverage the problem of large search spaces commonly affecting large-scale proteomics analyses [25–29]. Although the utility of a complete search is limited in practice, most information for calculating accurate posterior probabilities resides at the right tail of the peptide-spectrum match distribution. As a result, the guaranteed availability of just those candidates might be enough to reap most of the benefits of a complete search and make the analysis practical, an important step for further investigation. Our analysis thus aimed to show the relevance of prior probabilities in detection, which also allowed us to obtain accurate posterior probabilities even when considering large search spaces.

The biggest limitation of our study is the analysis of peptides from a single combinatorial peptide library—the analyzed peptides were all of the same sequence pattern. For instance, even though FFPSM$^{\star}_{\mathrm{P}}$ outperformed other *de novo* sequencing algorithms on this dataset, its general performance remains to be further evaluated. As FFPSM$^{\star}_{\mathrm{P}}$ requires the *a priori* distribution of fragment masses for all theoretical peptides at a given precursor mass range, making such comparison is harder in practice outside of a complete search strategy. Nevertheless, we consider this study as an illustration of the utility of a complete search for the development of peptide detection methods. Although we considered a prior model based on the distance to a single reference sequence which was of a limited utility (Fig. 6a and c), a similar prior based on the distance to *any* reference protein sequence is likely to allow reference-guided detection in a typical shotgun proteomics experiment. Although investigations along these lines were done earlier [20,21], direct integration of prior probabilities based on the distance to the reference sequence(s) into scoring is, to our best knowledge, missing. Notably, modeling of prior probabilities concerns the behavior of *peptides*, and thus such probabilities can be in principle incorporated into any detection method. The use of prior probabilities based on the expected presence of peptides in the sample will be thus most likely universally beneficial.

We did not consider Isoleucines in the analysis to allow for correctly interpreted spectra based purely on the spectral match. Otherwise, there will always be an additional match of an equal score as the correct peptide because the LVVVGA-**XX**-VGK pattern contains Leucine (e.g., if the correct peptide was **L**VVVGAGGVGK, the candidate **I**VVVGAGGVGK would have the same score in the analyzed metrics). Nevertheless, we would still obtain correctly interpreted spectra even if considering Isoleucines—if we utilized probabilistic prior knowledge. For instance, the sequence **L** → **I**VVVGAGGVGK is of distance one and thus less likely, providing discrimination between the sequences. Thus, utilizing probabilistic prior knowledge would also enable us to distinguish between Leucines and Isoleucines based on their prior probabilities, and in effect, posterior probabilities.

The complete search strategy can be thus considered as a well-defined environment for the development of peptide detection methods. The focus on complete search also allowed us to temporarily disregard various computational aspects of the analysis and study the pure potential of peptide-scoring metrics. The problems associated with large search spaces were largely suppressed by utilizing discriminative prior models, allowing detection against millions of candidates per spectrum. Integration of complete search and probabilistic prior knowledge thus allowed reliable peptide detection in a low-precursor-mass combinatorial peptide library, even when considering all theoretical peptides.

4. Methods

4.1. Synthetic peptide library

400 unpurified peptides of sequence LVVVGA-XX-VGK (XX being any combination of two coded amino acids) were ordered from JPT Peptide Technologies (Berlin, Germany). Peptides were analyzed individually on an LTQ Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) coupled to a Dionex UltiMate 3000 RSLCnano system (Dionex, Olten, Switzerland) via a Nanospray Flex Ion Source (Thermo Fisher Scientific, Bremen, Germany). For each LC-MS/MS run, 1 μL of a peptide (2 pmol) was injected. Peptides were loaded in buffer A (0.1% formic acid in water) and eluted from a 2 cm column (Acclaim PepMap 100, C18, 5 μm, 100 Å; Thermo Fisher Scientific, Bremen, Germany) using a linear 7.5-min gradient of 2%–40% of buffer B (0.1% formic acid in

acetonitrile) at a flow rate of 1 μL per minute. In each measurement cycle, a full MS scan was acquired using the Orbitrap analyzer (*mz* range 300−1700, 120K resolution). The 12 most abundant ions of intensity at least 10^3 were isolated (width ±1 *mz*), fragmented using CID (normalized collision energy: 35%), and measured in the ion trap (AGC target ion count of 10^4, 100 ms accumulation time). Already selected ions were excluded from repeated measurements for 30 s.

4.2. Candidate peptides

For each library peptide, we constructed all amino acid compositions within 10 ppm of the peptide's theoretical precursor mass (in-house backtracking script, carbamidomethylated cysteines). Afterward, we generated all candidate peptide sequences from the constructed amino acid compositions.

4.3. Scoring metrics

Although we evaluated the peptide-spectrum matches with the help of fragment indexation, herein, we express the scoring directly for simplicity. Suppose an observed spectrum with fragment masses $\mathbf{m}$ (m/z) and intensities $\mathbf{i}$. Now consider a predicted theoretical spectrum (for a particular peptide) with m/z $\mathbf{t}$. Suppose we match these spectra up to tolerance δ, thereby obtaining indices of matching peaks:

$$\mathbf{I}^{\mathbf{m}} = \langle I_1^m, \ldots, I_k^m \rangle,$$

$$\mathbf{I}^{\mathbf{t}} = \langle I_1^t, \ldots, I_k^t \rangle.$$

Thus for each $i \in \{1, \ldots, k\}$,

$$\left| \mathbf{m}_{I_k^m} - \mathbf{t}_{I_k^t} \right| \leq \delta.$$

Further, if an experimental peak matches multiple theoretical peaks, we retain only the first such index in $\mathbf{I}^t$ and a corresponding index in $\mathbf{I}^m$. We now continue with the description of the individual scoring metrics.

4.3.1. Number of matching peaks (NMP)

The number of matching peaks is simply the k of matching peaks.

4.3.2. Rank-normalized sum of intensities (RPSM)

First, the metric replaces the intensity vector $\mathbf{i}$ with its ranks. The metric then sums the intensities of matching fragments, thus:

$$\sum_{i=1}^{k} \mathbf{i}_{I_k^m}.$$

4.3.3. Hyperscore (HSPSM)

Analogously as for RPSM, the intensity vector $\mathbf{i}$ is replaced with its ranks. Let b represent the number of matching b ions of a candidate peptide (as in NMP), and analogously let y represent the number of matching y ions. The HSPSM is then defined as:

$$\log_{10}\left(\sum_{i=1}^{k} \mathbf{i}_{I_k^m} \cdot b! \cdot y! \right)$$

The HSPSM can be thus thought of as an extension of RPSM.

4.3.4. Fragment frequency (FFPSM)

The metric exploits the distribution of predicted fragment masses of all candidate peptides for a spectrum. As theoretical spectra for many candidate peptides often share the same fragment masses, such fragments are more likely *a priori*. And although rare fragments increase the ability to match a particular peptide uniquely, their low frequency also makes them prone to match noise peaks. In essence, FFPSM behaves such that the less likely a fragment mass is *a priori*, the greater intensity it requires to be considered relevant. Formally, suppose all theoretical candidate peptides $\mathbf{p} = \langle p_1, \ldots, p_n \rangle$ for a spectrum. For each $i \in \{1, \ldots, n\} = \mathbb{I}$, suppose a theoretical spectrum $\mathbf{t}^i$ containing all its predicted fragments. For each mass of a theoretical fragment m, let $f(m)$ be the number of theoretical spectra that contain it, thus:

$$f(m) = \left| \left\{ i \in \mathbb{I} \;\middle|\; m \in \mathbf{t}^i \right\} \right|.$$

The match is then defined as:

$$\sum_{i=1}^{k} \log_{10}\left(\mathbf{i}_{I_k^m} \cdot f\left(\mathbf{t}_{I_k^m} \right) \right).$$

4.3.5. Length-based score adjustment

We utilized the following adjustment to scoring to suppress the correlations of scores and peptide lengths. Suppose a vector of scores

$$\mathbf{m} = \langle \mathbf{m}_1, \ldots, \mathbf{m}_n \rangle,$$

such that each $\mathbf{m}_i$ corresponds element-wise to a peptide p_i of length l_i, $i \in \{1, \ldots, n\}$, where n is the total number of candidate peptides for a spectrum. For each peptide length l, denote $\mathbf{m}^l$ the subvector of $\mathbf{m}$ that contains just the peptides of length l. We then defined the length-adjusted score $\mathbf{m}_i^{\star}$ as

$$\mathbf{m}_i^{\star} = \mathbf{m}_i - \overline{\mathbf{m}^{l_i}}$$

where $\overline{e}$ represents the average value of e.

4.4. Calculation of posterior probabilities

4.4.1. Model

Ideally, we would like to know the probability that a peptide p produced a spectrum s, $\Pr(p \mid s)$. However, calculating such probability might require $\Pr(s \mid p)$, the probability of observing a given spectrum for a peptide p, which can be complicated. Instead of the spectrum itself, we work with a vector $\mathbf{m}$ of peptide-spectrum matches of all candidate peptides with the spectrum. Directly, we would work with the whole match vector $\mathbf{m}$, thus calculating $\Pr(p \mid \mathbf{m})$, the probability of peptide p after observing a match vector $\mathbf{m}$. Nevertheless, we use a much simpler model: the probability of a peptide p given its match x, thus $\Pr(p \mid \mathbf{m}_p = x)$. We use Bayes' Theorem to derive the probability, giving

$$\Pr(p \mid \mathbf{m}_p = x) = \frac{\Pr(\mathbf{m}_p = x \mid p) \cdot \Pr(p)}{\Pr(\mathbf{m}_p = x)}. \tag{1}$$

4.4.2. Assumptions

For further simplification, we assume that observing a particular match for a peptide—assuming the peptide is true—is independent of the peptide. Thus, for any match x,

$$\Pr(\mathbf{m}_p = x \mid p) = \Pr(\mathbf{m}_q = x \mid q). \tag{2}$$

Such an assumption allows us to learn the behavior of correct

matches from all peptides in the training dataset. Similarly, we assume that the probability of observing a particular match at random is independent of the peptide. Thus, for any match x,

$$\Pr(\mathbf{m}_p = x) = \Pr(\mathbf{m}_q = x). \tag{3}$$

4.4.3. Correct matches

Suppose a dataset $\mathbb{D}$ of match vectors and corresponding correct peptides,

$$\mathbb{D} = \left\{\langle \mathbf{m}^1, q_1\rangle, \ldots, \langle \mathbf{m}^n, q_n\rangle\right\},$$

and an indexing set $\mathbb{I} = \{1,\ldots,n\}$. As we assume the independence by (2), we set the probability of any peptide p having a match x as the overall proportion at which correct peptides have a match x,

$$\Pr(\mathbf{m}_p = x \mid p) = \frac{\left|\left\{i \in \mathbb{I} \,\middle|\, \mathbf{m}^i_{q_i} = x\right\}\right|}{|\mathbb{I}|}.$$

4.4.4. Random matches

As we assume the independence by (3), we set the probability of $\mathbf{m}_p = x$ as the overall proportion of x in all match vectors from $\mathbb{D}$. Thus, suppose a vector $\mathbf{M}$ that is a concatenation of all match vectors $\mathbf{m}^i$ for $i \in \mathbb{I}$. Let its total length be l, and its corresponding indexing set $\mathbb{J} = \{1,\ldots,l\}$. We set

$$\Pr(\mathbf{m}_p = x) = \frac{\left|\{j \in \mathbb{J} \mid \mathbf{M}_j = x\}\right|}{|\mathbb{J}|}.$$

4.4.5. One-sum normalization

Due to our approximations, the posterior probabilities calculated using (1) do not generally sum to one. We neglect such imprecision and because we analyze the behavior of a complete search, we normalize the posterior probabilities to sum to one.

4.4.6. Calculation of NMP_p

Instead of raw peptide-spectrum matches, we transformed them into their *relative* form—the distance to the best-matching candidate (per spectrum). Thus, the best-matching candidates have a relative fragment match of 0, those with one matching peak less have a relative match of 1, and so forth. Note that the distribution of relative fragment matches resembled a geometric distribution, with its only parameter r set to the proportion of spectra in which the correct peptide had a maximal match (Supplementary Fig. 1). Although the use of a fixed true match distribution worked already reasonably well (data not shown), we parameterized the true match distribution to account for further structure in the data. For instance, the proportion of spectra in which a correct peptide had a maximal match increased moderately with the intensity of precursors (Spearman's $\rho = 0.30$, $p = 7.78 \times 10^{-46}$, $n = 2\,144$), and we wanted to take such and other dependence into account. In particular, we predicted the probability r that a peptide has a maximal match based on spectral characteristics and constructed a geometric distribution of true matches with r as its only parameter. We used logistic regression for the prediction of r, utilizing precursor intensity, precursor mass, and the total number of candidate peptides as independent variables (LogisticRegression from sklearn, $\log_{10}$ of independent variables). For the distribution of random matches $\Pr(\mathbf{m}_p)$, we just directly summed the distributions of matches over all spectra (Supplementary Fig. 2).

4.4.7. Calculation of $FFPSM^{\star}_{p}$

Similarly as for NMP_p, we transformed the matches into their relative form. As the behavior $FFPSM^{\star}_{p}$ was by far less discrete than NMP_p, we fit an exponential distribution over the distribution of true matches (expon from scipy.stats). Note that the exponential distribution did not fit sufficiently well the behavior of true matches, partially because of aggregation of correct matches at a relative maximum match around 5 (Supplementary Fig. 3). Such behavior resulted in the underestimation of probabilities of best candidates and partially worsened the performance of $FFPSM^{\star}_{p}$; however, we left an examination of the reasons for future investigations. For the behavior of random matches, we merged the distributions of matches from all spectra and fit a gaussian distribution over them (norm from scipy.stats, Supplementary Fig. 4). Note that as the norm function did not allow fitting data expressed as distributions (as values and their counts), we sampled 10^6 relative maximum matches and fit the gaussian distribution over them.

4.5. Comparison with existing software

4.5.1. Common configuration

Each algorithm was run with the following parameters: precursor tolerance (10 ppm), fragment tolerance (0.3), carbamidomethylation of cysteines set as a fixed modification, and no variable modifications allowed. Each algorithm was run with MS/MS spectra in mgf format, and the fragment ions were restricted to the 100 most-intense peaks. *De novo* sequencing algorithms were run with the trypsin enzyme to allow comparison to the performance of detection when prior probabilities of candidates followed the expected cleavage behavior. Database search algorithms were run with no cleavage in order to compare the evaluation of their spectral scoring metrics to the analyzed scoring metrics.

4.5.2. Detailed configuration

Novor (v1.06.063) was run with CID for fragmentation and Trap for mass analyzer. PepNovo (release 20101117) was run with CID_IT_TRYP model, parameters -use_spectrum_charge, -use_spectrum_mz, -num_solutions 2000, -no_quality_filter. As PepNovo did not support precursor tolerance in ppm, we ran it separately for each peptide, with the parameter -pm_tolerance set to the absolute mass tolerance corresponding to the tolerance of 10 ppm for the mass of the correct peptide. We used DeepNovo (v. 0.0.1) from the master branch at https://github.com/nh2tran/DeepNovo accessed on 2020/11/26, and downloaded the pretrained model train.-example as specified in the README. Following the example in the README, we used the parameters −−beam_search and −−beam_size 5 for sequencing *de novo*. Comet (v. 2020.01 rev. 4) was run with the following parameters: no isotope errors (isotope_error = 0), no minimal number of peaks in spectrum (minimum_peaks = 0), I and L treated as different residues (equal_I_and_L = 0), and a newly specified enzyme "Cut_nowhere" (11. Cut_nowhere 0 I I) designed to avoid cleavage (no Isoleucines in the database), and setting of the corresponding enzyme (sample_enzyme = 11). For the analysis using single-charged fragments, we set max_fragment_charge = 1, otherwise we kept the default max_fragment_charge = 3. MS-GF+ (v. 20210322) was run with the following parameters: no isotope error (-ti 0,0), no cleavage (-e 9), fully-specific peptides (-ntt 2), standard protocol (-protocol 5), and 100 best candidates per spectrum (-n 100). MSFragger (v. 3.2) was run using closed-search parameters (closed_fragger.params), with further changes: no isotope error

(isotope_error = 0), no mass calibration (calibrate_mass = 0), no cleavage (search_enzyme_cutafter = U), no deisotoping (deisotope = 0), no removal of neutral losses (deneutralloss = 0), singly-charged fragments (max_fragment_charge = 1), maximal allowed number of missed cleavages (allowed_missed_cleavage = 5), no variable mods (max_variable_mods_per_peptide = 0), 100 reported results per spectrum (output_report_topN = 100), high maximal E-value (output_max_expect = 5000000), lower minimal digest length (digest_min_length = 6), no minimal fragment intensity ratio (minimum_ratio = 0). Similarly as for the Comet, for the analysis using single-charged fragments we set max_fragment_charge = 1, otherwise we kept the default max_fragment_charge = 3. Note that we set the configuration of MSFragger to obtain close correspondence of its scoring metric to the HSPSM.

Data availability

The data for the peptide library have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (ID: PXD013421).

CRediT authorship contribution statement

Miroslav Hruska: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Dusan Holub:** Investigation, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Miroslav Hruska reports a relationship with Ministry of Education, Youth and Sports of the Czech Republic that includes: funding grants. Miroslav Hruska reports a relationship with Technology Agency of the Czech Republic that includes: funding grants. Miroslav Hruska reports a relationship with Ministry of Health of the Czech Republic that includes: funding grants. Miroslav Hruska reports a relationship with Horizon 2020 that includes: funding grants. Dusan Holub reports a relationship with Ministry of Education, Youth and Sports of the Czech Republic that includes: funding grants. Dusan Holub reports a relationship with Technology Agency of the Czech Republic that includes: funding grants. Dusan Holub reports a relationship with Ministry of Health of the Czech Republic that includes: funding grants. Miroslav Hruska has patent #Method of identification of entities from mass spectra (PCT/EP2019/069552) pending to Univerzita Palackeho v Olomouci.

Acknowledgement

This work was supported in parts by the Ministry of Education, Youth and Sports of the Czech Republic (CZ.02.1.01/0.0/0.0/16_019/0000868, CZ.01.1.02/0.0/0.0/16_084/0010360, LM2015064, LM2015047, LM2018130, LM2018131), Technology Agency of the Czech Republic (TE02000058, TN01000013), Ministry of Health of the Czech Republic (NV16-32318A, NV16-32302A), and the European Union's Horizon 2020 (EOSC-Life Grant agreement no. 824087).

Appendix A. Supplementary data

The DOI of the dataset is https://doi.org/10.17632/3j95c7tm5t.1.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijms.2021.116723.

References

[1] R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function, Nature 537 (2016) 347–355.
[2] M. Wilhelm, J. Schlegl, H. Hahne, A.M. Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, Mass-spectrometry-based draft of the human proteome, Nature 509 (2014) 582–587.
[3] K. Verheggen, H. Raeder, F.S. Berven, L. Martens, H. Barsnes, M. Vaudel, Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows, Mass Spectrom. Rev. (2017) 1–15.
[4] J.K. Eng, A.L. Mccormack, J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, American society for Mass Spectrometry 5 (1994) 976–989.
[5] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (2004) 1466–1467.
[6] S. Kim, P.A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics, Nat. Commun. 5 (2014).
[7] B. Wen, W.F. Zeng, Y. Liao, Z. Shi, S.R. Savage, W. Jiang, B. Zhang, Deep Learning in Proteomics, 2020. Proteomics.
[8] X.X. Zhou, W.F. Zeng, H. Chi, C. Luo, C. Liu, J. Zhan, S.M. He, Z. Zhang, PDeep: predicting MS/MS spectra of peptides with deep learning, Anal. Chem. 89 (2017) 12690–12697.
[9] W.F. Zeng, X.X. Zhou, W.J. Zhou, H. Chi, J. Zhan, S.M. He, MS/MS Spectrum prediction for modified peptides using pDeep2 Trained by Transfer Learning, Anal. Chem. 91 (2019) 9724–9731.
[10] K. Liu, S. Li, L. Wang, Y. Ye, H. Tang, Full-spectrum prediction of peptides tandem mass spectra using deep neural network, Anal. Chem. 92 (2020) 4275–4283.
[11] J.M. Chick, D. Kolippakkam, D.P. Nusinow, B. Zhai, R. Rad, E.L. Huttlin, S.P. Gygi, A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides, Nat. Biotechnol. 33 (2015) 743–749.
[12] O.S. Skinner, N.L. Kelleher, Illuminating the dark matter of shotgun proteomics, Nat. Biotechnol. 33 (2015) 717–718.
[13] A.T. Kong, F.V. Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics, Nat. Methods 14 (2017) 513–520.
[14] A. Devabhaktuni, S. Lin, L. Zhang, K. Swaminathan, C.G. Gonzalez, N. Olsson, S.M. Pearlman, K. Rawson, J.E. Elias, TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets, Nat. Biotechnol. 37 (2019) 469–479.
[15] T. Muth, F. Hartkopf, M. Vaudel, B.Y. Renard, A potential golden age to come—current tools, recent use cases, and future avenues for de novo sequencing in proteomics, Proteomics 18 (2018).
[16] A. Frank, P. Pevzner, PepNovo: de novo peptide sequencing via probabilistic network modeling, Anal. Chem. 77 (2005) 964–973.
[17] B. Ma, Novor: real-time peptide de Novo sequencing software, J. Am. Soc. Mass Spectrom. 26 (2015) 1885–1894.
[18] N.H. Tran, X. Zhang, L. Xin, B. Shan, M. Li, De novo peptide sequencing by deep learning, Proc. Natl. Acad. Sci. U. S. A 114 (2017) 8247–8252.
[19] H. Yang, H. Chi, W.F. Zeng, W.J. Zhou, S.M. He, PNovo 3: precise de novo peptide sequencing using a learning-to-rank framework, Bioinformatics 35 (2019) i183–i190.
[20] I.V. Shilov, S.L. Seymour, A.A. Patel, A. Loboda, W.H. Tang, S.P. Keating, C.L. Hunter, L.M. Nuwaysir, D.A. Schaeffer, The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra, Mol. Cell. Proteomics : MCP 6 (2007) 1638–1655.
[21] B.Y. Renard, B. Xu, M. Kirchner, F. Zickmann, D. Winter, S. Korten, N.W. Brattig, A. Tzur, F.A. Hamprecht, H. Steen, Overcoming species boundaries in peptide identification with bayesian information criterion-driven error-tolerant peptide search (BICEPS), Mol. Cell. Proteomics 11 (2012). M111.014167–1–M111.014167–12.
[22] N. Zhang, R. Aebersold, B. Schwikowski, ProbID, A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data, Proteomics 2 (2002) 1406–1412.
[23] Z. An, L. Zhai, W. Ying, X. Qian, F. Gong, M. Tan, Y. Fu, PTMiner: localization and quality control of protein modifications detected in an open search and its application to comprehensive post-translational modification characterization in human proteome, Mol. Cell. Proteomics 18 (2019) 391–405.
[24] D.D. Shteynberg, E.W. Deutsch, D.S. Campbell, M.R. Hoopmann, U. Kusebauch, D. Lee, L. Mendoza, M.K. Midha, Z. Sun, A.D. Whetton, R.L. Moritz, PTMProphet: fast and accurate mass modification localization for the trans-proteomic pipeline, J. Proteome Res. 18 (2019) 4262–4272.
[25] A.I. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat. Methods 11 (2014) 1114–1125.
[26] W.S. Noble, Mass spectrometrists should search only for peptides they care about, Nat. Methods 12 (2015) 605–608.
[27] M.V. Ivanov, A.A. Lobas, D.S. Karpov, S.A. Moshkovskii, M.V. Gorshkov, Comparison of false discovery rate control strategies for variant peptide identifications in shotgun proteogenomics, J. Proteome Res. 16 (2017) 1936–1943.

[28] K. Li, A. Jain, A. Malovannaya, B. Wen, B. Zhang, DeepRescore: leveraging deep learning to improve peptide identification in immunopeptidomics, Proteomics 20 (2020) 1–10.
[29] J.A. Vizcaíno, P. Kubiniok, K.A. Kovalchik, Q. Ma, J.D. Duquette, I. Mongrain, E.W. Deutsch, B. Peters, A. Sette, I. Sirois, E. Caron, The human immunopeptidome project: a roadmap to predict and treat immune diseases, Mol. Cell. Proteomics 19 (2020) 31–49.
[30] K.L. Bryant, J.D. Mancias, A.C. Kimmelman, C.J. Der, KRAS: feeding pancreatic cancer proliferation, Trends Biochem. Sci. 39 (2014) 91–100.
[31] J.K. Eng, M.R. Hoopmann, T.A. Jahan, J.D. Egertson, W.S. Noble, M.J. MacCoss, A deeper look into Comet - implementation and features, J. Am. Soc. Mass Spectrom. 26 (2015) 1865–1874.
[32] D.L. Tabb, Y. Huang, V.H. Wysocki, J.R. Yates, Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides, Anal. Chem. 76 (2004) 1243–1248.
[33] S.L. Hubler, P. Kumar, S. Mehta, C. Easterly, J.E. Johnson, P.D. Jagtap, T.J. Griffin, Challenges in peptide-spectrum matching: a robust and reproducible statistical framework for removing low-accuracy, high-scoring hits, J. Proteome Res. 19 (2020) 161–173.
[34] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat. Biotechnol. 26 (2008) 1367–1372.
[35] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat. Methods 4 (2007) 207–214.
[36] L. Käll, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, Nat. Methods 4 (2007) 923–925.