# SpecPeptidOMS Directly and Rapidly Aligns Mass Spectra on Whole Proteomes and Identifies Peptides That Are Not Necessarily Tryptic : Implications for Peptidomics

Émile Benoist, Géraldine Jean, Hélène Rogniaux, Guillaume Fertin, Dominique Tessier

Journal of proteome research

pubs.acs.org/jpr

Technical Note

# *SpecPeptidOMS* Directly and Rapidly Aligns Mass Spectra on Whole Proteomes and Identifies Peptides That Are Not Necessarily Tryptic: Implications for Peptidomics

*Published as part of Journal of Proteome Research special issue "Software Tools and Resources 2025".*

Émile Benoist, Géraldine Jean,* Hélène Rogniaux, Guillaume Fertin, and Dominique Tessier

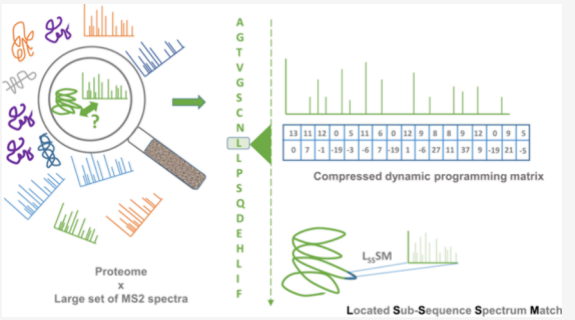
Cite This: *J. Proteome Res.* 2025, 24, 2159−2172

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



**ABSTRACT:** *SpecPeptidOMS* directly aligns peptide fragmentation spectra to whole and undigested protein sequences. The algorithm was specifically and initially designed for peptidomics, where the aim is to identify peptides that do not result from the hydrolysis of a known protein and therefore, whose termini cannot be predicted. Thus, *SpecPeptidOMS* can perform alignments starting and ending anywhere in the protein sequence. The underlying computational method of *SpecPeptidOMS*, which is based on a dynamic programming approach, was drastically optimized. As a result, *SpecPeptidOMS* can process around 12,000 spectra per hour on an ordinary laptop, with alignment performed against the entire human proteome. The performance of *SpecPeptidOMS* was first evaluated on a publicly available data set of (nontryptic) synthetic mass spectra. Accuracy was estimated by considering the results obtained by *MaxQuant* on the same data set as the "ground truth". A second series of tests on a larger, well-known proteomics data set (HEK293) highlighted *SpecPeptidOMS*' additional ability to search for open modifications, a feature of interest in peptidomics but also more broadly in conventional proteomics. *SpecPeptidOMS* is open-source, cross-platform (written in Java), and freely available.




**KEYWORDS:** *Peptidomics, proteomics, mass spectrometry, MS2 spectra, dynamic programming*

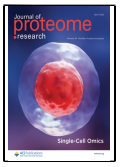
## INTRODUCTION

In bottom-up proteomics, the most common approach to deducing a peptide from an experimental fragmentation mass spectrum (MS2 spectra) relies on the best fit with one of the theoretical models to which it is compared. Those theoretical models are obtained by simulating ideal fragmentation from theoretical peptides generated from protein databases using a sequence cleavage model specified according to the digestion enzyme used in the wet lab experiment. To avoid a prohibitive number of comparisons, conventional algorithms restrict the search space by limiting the comparison of experimental spectra to the theoretical model spectra that fall within a narrow mass window.[1] This computational approach enables the rapid and effective identification of spectra that can be validated by consensual statistics. However, spectra that deviate from the model in a nonanticipated way, for example because they carry unexpected chemical modifications (post-translational modifications or PTMs)[2] or result from protein cleavage at unanticipated site, will remain unidentified[3] when not contributing to incorrect interpretations.[4] Strikingly, these two situations are very common, and indeed almost inherent, in the context of peptidomics.

Peptidomics is an emerging field of the omics sciences,[5−8] that refers to the identification of the pool of peptides present in a biological fluid or tissue. The field has grown rapidly in the past few years because endogenous peptides have been established as essential players in cellular processes (e.g., signaling, immune response, intercellular communication, homeostasis, etc.) and potential biomarkers for a number of cellular disorders.[9] Nonendogenous peptides are another group of circulating peptides that notably include food-derived peptides, and food peptidomics is arousing great interest.[10]

From a bioinformatics perspective, interpreting MS2 spectra in peptidomics adds complexity[11] compared to proteomics because no assumption can be made about cleavage sites, meaning that the peptide ends can lie at any amino acid of the protein. Another characteristic of peptidomics is that peptides

**Received:** September 30, 2024
**Revised:** January 17, 2025
**Accepted:** February 25, 2025
**Published:** March 27, 2025



can be highly modified: this characteristic is almost inherent to their bioactivity, as modifications protect them from overly rapid proteolysis in biological fluids. Today, biologists conducting peptidomics studies choose among several options for interpreting MS2 spectra. The first assumes that the peptides of interest have distinctive characteristics, enabling the search to be targeted, and the search space to be reduced. In a second option, the spectra are interpreted using a conventional algorithm (i.e., designed for proteomics), with a nonspecific digestion mode taking into account all possible peptides between a minimum and maximum length along the proteins. The latter option results in huge search spaces and realistically eliminates the possibility of considering many types of modifications in addition to the absence of a cleavage rule. A third option is to use search engines that incorporate a step of *de novo* sequencing, such as PEAKS (Bioinformatics Solutions Inc.),[12] which has proved effective in the context of peptidomics.[13] Yet, *de novo* results still require a manual inspection that is hardly compatible with large-scale experiments. In summary, although the field is booming in several area of life sciences, there is currently no approach for peptide identification that has been designed specifically for peptidomics. The manuscript introduces *SpecPeptidOMS*, a software that has been originally developed to address the challenge of peptide identification in peptidomics.

*SpecPeptidOMS* is a major upgrade of the *SpecGlobX* algorithm, which our group has recently proposed for the identification of multiple and unbiased modifications of tryptic peptides.[14] Both algorithms fall in the family of the so-called "open modification search" (OMS) methods that have emerged over the past decade in proteomics. OMS methods have implemented advanced computational optimization techniques accelerating spectra comparison to a sufficient extent to widen the mass window of spectra comparison. As a result, those methods can identify peptides carrying unanticipated modifications. However, if the identification and localization of a single modification are successful, the presence of multiple modifications in a mass spectrum remains problematic. *SpecGlobX*, based on an efficient dynamic programming algorithm, can align pairs of spectra quickly while detecting several modifications.[14] To limit execution time, *SpecGlobX* runs on a set of peptide sequence matches (PSMs) generated by another OMS search engine (*SpecOMS*,[15] for example). Compared to SpecGlobX, *SpecPeptidOMS*, incorporates two fundamental design changes that deeply impact its capabilities. First, a new and condensed representation of experimental spectra allows direct alignment of MS2 spectra to undigested proteins without preconceptions about where the alignments should start and end in the protein. Second, a drastic optimization boosts the execution time by several orders of magnitude while keeping the memory requirements low. Importantly, *SpecPeptidOMS* retains the advantage inherited from *SpecGlobX* in identifying peptides carrying multiple modifications that had not been anticipated. However, while the quality of the *SpecGlobX* results were depending on the relevance of the set of PSMs used as input, *SpecPeptidOMS* is independent of any other tool and evaluates all possible PSMs. Then, the new design of *SpecPeptidOMS* has opened the way for efficient identification of MS2 spectra arising from peptidomics data sets.

In the present paper, we explain how the drastic optimization of the dynamic programming approach behind *SpecPeptidOMS* allowed handling direct alignment of approximately 12,500 spectra per hour on a human proteome at a laptop pace. Next, to ensure that *SpecPeptidOMS* meets the needs of peptidomics spectra identification in terms of sensitivity and accuracy, we compare its results to those provided by *Andromeda*,[16] the search engine integrated into the *MaxQuant* environment, on a spectra data set arising from synthetic (nontryptic) peptides. Next, we run *SpecPeptidOMS* on a larger proteomics data set to evaluate its behavior in the presence of various PTMs and chimeric spectra. We discuss the strengths and weaknesses of the *SpecPeptidOMS* approach, based on selected examples.

Although this paper can be read in itself, some essential concepts are only briefly described considering that they are already detailed elsewhere.[14]

## MATERIAL AND METHODS

### Data sets and Protein Database

Several spectra data sets were downloaded from PRIDE[17] to evaluate *SpecPeptidOMS*. The first data set corresponds to 941 nontryptic synthetic peptides representing HLA class II ligands generated in the ProteomeTools project (PXD021013, file 03035a_GA3-TUM_HLA2_3_01_01−3xHCD-1h-R1.raw).[18] This spectra data set contains 63,873 spectra acquired on an orbitrap using an inclusion list of precursors, corresponding to the expected peptides in several charge states. After conversion into the MGF Format with msConvert release 3.0,[19] this data set referred to as *Synthetic_HLA2* for short in the manuscript contains 59,460 MS2 spectra.

We downloaded the corresponding interpretation of spectra run using *MaxQuant* against a specific fasta file corresponding to an extended version of the 941 synthetic peptides. The results were filtered at 1% peptide FDR (file msms_db.csv downloaded) and finally provided 40,331 interpretations. The second data set (PXD001468) comprises 24 spectra data sets generated from HEK293 cells.[20] Sample preparation and spectra acquisition details can be found in the original publications for both data sets.[18,20,21]

The human database used in all the tests reported in the manuscript was downloaded from UniProt,[22] with one protein sequence per gene and contains 20,591 proteins (UP000005640_9606.fasta). The database of synthetic peptide sequences was obtained by flanking the sequences of the synthetic peptides identified by *MaxQuant* by their N-Term and C-Term cleavage window sequences (file peptides.txt downloaded). This process generates peptides with lengths ranging from 23 to 33 amino acids whose central parts are synthetic peptides.

### Experimental Mass Spectra and Protein Data Preprocessing

Before alignment, a rapid pretreatment (less than a minute for about 57,000 spectra) prepares spectra data structures. Because no information on the boundaries of fragmented peptides is provided, *SpecPeptidOMS* aligns experimental spectra on protein amino acids without generating theoretical spectra. By convention, *SpecPeptidOMS* represents each protein $P$ by its reverse amino acid sequence $P^R$ (from the C-terminal to the N-terminal extremities). In fact, *SpecPeptidOMS* makes the arbitrary choice to consider fragments in the spectrum as y-ions, based on the fact that y-ions (containing the C-terminus of the peptide) are generally more intense in the spectrum than b-ions.[23] Thus, to ensure that the spectrum (in y-ions) is read


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

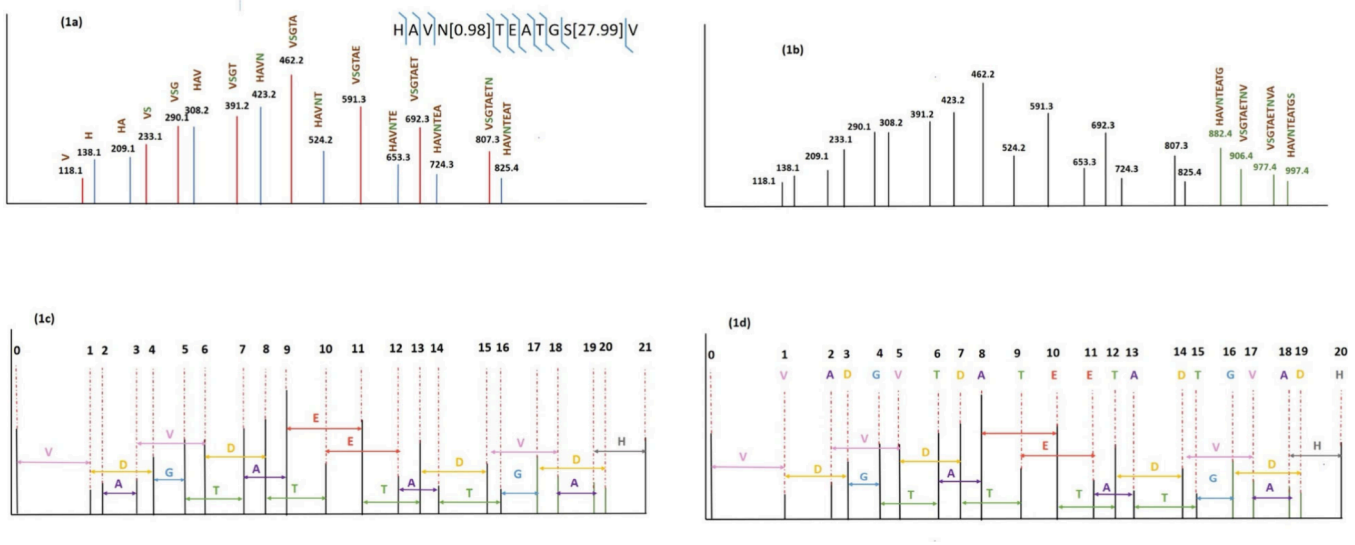



**Figure 1.** Preprocessing of the simulated spectrum Se1. (1a) *Se1* is a simulated experimental spectrum representing the peptide HAVN[0.98]TEATGS[27.99]V containing two modifications: deamidation of residue N and formylation of residue G. The blue peaks (red peaks) correspond to b-ions (respectively y-ions). The intensity of the peaks is only represented to mimic an experimental spectrum, although it is not used by the algorithm.; (1b) $Se_c1$ is the completed spectrum obtained from *Se1*, comprising original peaks (in black) plus added complementary peaks (in green); (1c) amino acids are labeled in $Se_c1$ and two peaks respectively representing the beginning and the end of the y-ions are added to the spectrum (peaks numbered 0 and 21); (1d) masses that are useless for the alignment are removed in $Se_c1$ and labeled amino acids are reindexed in $Se_c1'$. For this example, the second mass index in $Se_c1$ is removed because this mass index is not the upper mass limit of any amino acid, so $Se_c1'$ contains 21 mass indexes instead of 22 mass indexes before the treatment. Finally, each labeled amino acid is represented by the mass index of its upper mass limit.

in the same direction as the protein sequence, the sequence was reversed into $P^R$.

Here we recall that if two peaks in an MS2 spectrum are such that their mass difference is equal to the mass of one amino acid, they can be used to label the amino acid that is flanked by these two peaks. Once performed, *SpecPeptidOMS* should ideally align these labeled peaks on the amino acids of $P^R$ via a dynamic programming approach.

In the first processing step, masses of every experimental spectrum *Se* are filtered out to keep only the *n* most intense, as defined by the user. Next, *SpecPeptidOMS* generates a new spectrum, which we call $Se_c$ (for the completed spectrum), by adding new fragment masses called "complementary masses" to *Se* as in Prunier et al.[14] The rationale behind this addition of masses is to complete the spectrum when y-ion peaks are missing, whereas their corresponding b-ion peaks are in *Se*. We illustrate these transformations in one example in Figures 1a and 1b. Then, *SpecPeptidOMS* enumerates all of the amino acids that can be labeled in $Se_c$ by computing mass differences between all pairs of peaks. The result is a compact data structure where every experimental spectrum *Se* is described jointly by the list of its masses and amino acid labels. Labeled amino acids are represented by their upper mass indexes in $Se_c$ (see Figure 1c). This dual representation is convenient because, if labeled peaks in *Se* are aligned on protein amino acids, masses are useful to introduce mass shifts in the alignments. After this step, all the masses that are not the upper mass limit of any amino acid are useless for alignments: they are removed from the spectra representation to gain speed, generating a new spectrum $Se_c'$. This last transformation is illustrated in Figure 1d, where each potential amino acid is reindexed to account for the mass that has been removed.

## Algorithm Overview (Part 1): Fast Preliminary Mass Spectra Alignments on Unprocessed Proteins to Determine Located Sub-Sequence Spectrum Match ($L_{SS}SM$)

*SpecPeptidOMS* aligns the spectra in two rounds. First, the preliminary alignment round associates each spectrum with the protein(s) it aligns the best, generating something similar to PSMs, except that the locations of subsequences of $P^R$ replace peptides ($L_{SS}SM$ for Located subsequence Spectrum Match).

Alignments are processed protein by protein. For each $Se_c'$, *SpecPeptidOMS* fills in a matrix *D* that scores the alignment of $Se_c'$ along the reverse protein $P^R$ to find the ordered list of labeled amino acids in $Se_c'$ that best aligns with $P^R$. In a first approximation, one can imagine *D* as a large two-dimensional matrix containing as many rows as amino acids in $P^R$ and as many columns as mass indexes in $Se_c'$. The cell *D[i,j]* scores the best alignment of the amino acid at position i in $P^R$ (called $aa_i$) on the $j^{th}$ mass index of $Se_c'$ via a dynamic programming approach. We illustrate briefly how D is filled in Figure 2 (remind that a detailed explanation of the algorithm is available elsewhere[14]).

Once the basics for computing the scores in the D matrix have been established, we can optimize this process. *SpecPeptidOMS* retrieves which mass index(es) can be aligned on $aa_i$ instantly thanks to the representation of $Se_c'$ as a structure that links each amino acid to the list of its position(s) in the spectrum. When $aa_i$ has no corresponding label in $Se_c'$, all the cells in the i-th row of *D* are penalized because $aa_i$ cannot be found. This situation is called *not-found* and is illustrated in Figure 3. Since the amino acids K, M, S, N and I are not labeled in $Se_c1'$, their rows in *D* are colored in gray, meaning that the scores of all these cells are deduced from their preceding row by applying the formula: *D[i,j]=D[i-1,j] - penalty* (whose value is set to −4 by default). When one or


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

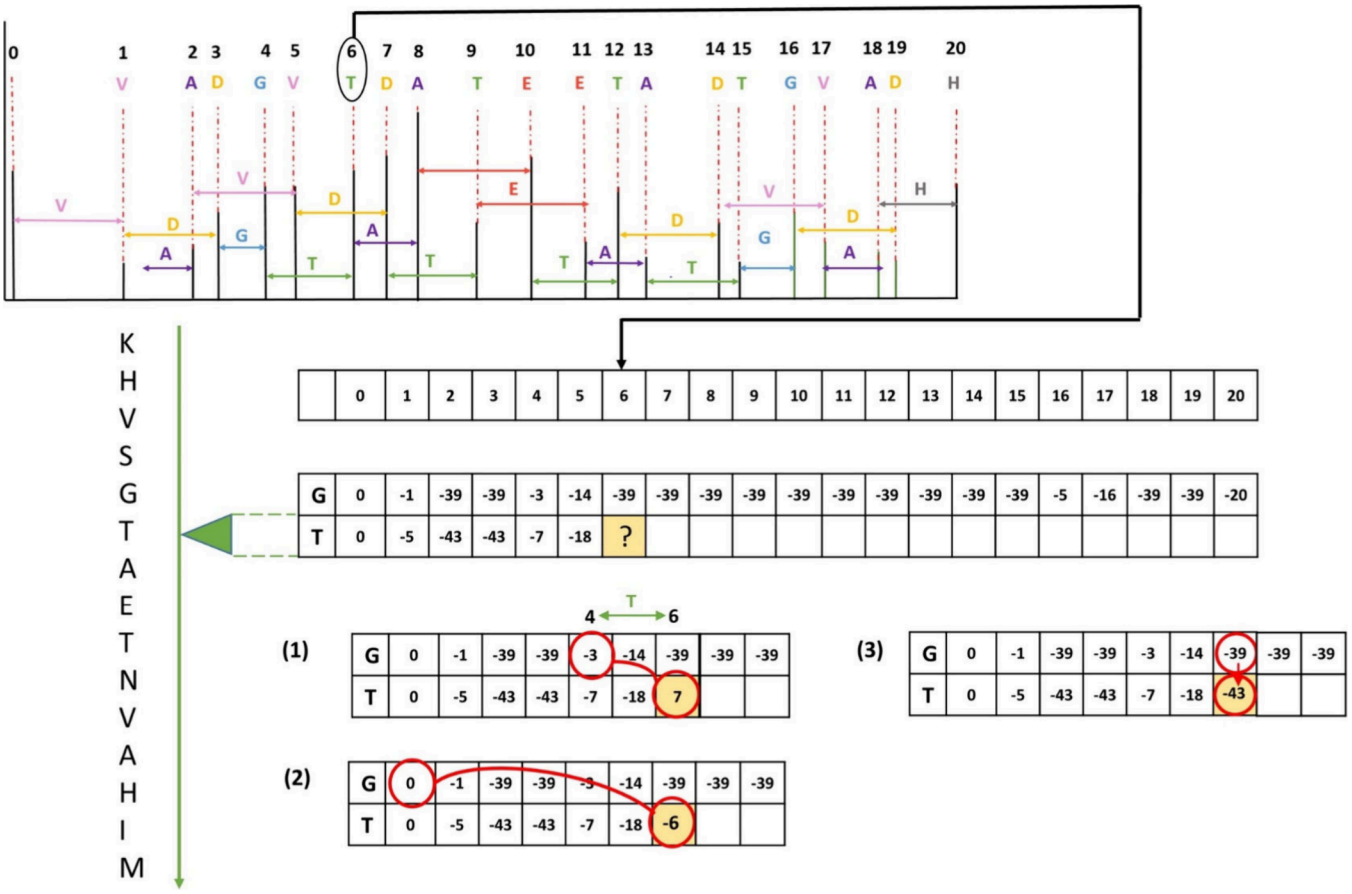



**Figure 2.** Alternatives were evaluated to fill a cell in the dynamic programming matrix D. *SpecPeptidOMS* fills D row by row, sliding amino acid by amino acid the fictitious reverse protein $P_1^R$ whose sequence is MI**HAVNTEATGSV**HK when spectrum *Se1*$_c$' representing peptide HAVN[0.98]TEATGS[27.99]V (from Figure 1) is aligned. Each row in D represents one amino acid, and only the mass indexes labeled with this amino acid can be aligned. Then, where applicable, SpecPeptidOMS evaluates three alternatives to align a mass index labeled with an amino acid on that amino acid in $P^R$. The schema presents the alignment of the 6th mass index of $Se_c1$' (labeled T) on the 6th amino acid of $P_1^R$*(i.e., T)*. In alternative (1), the 6th mass index is aligned without any shift (continuing the alignment done for the 5th amino acid (G) at the 4th mass index, with the 4th mass index being the lower bound mass of T). Then, D[6,6] = D[5,4] + 10 = 7, if 10 is the bonus attributed to an alignment without shift; in alternative (2), *SpecPeptidOMS* evaluates the possibility of a shift from the mass index lower than 6 having the greatest score on the preceding row (here, the mass index 0). Then, D[6,6] = D[0,5] - 6 = −6 if −6 is the penalty associated with a shift; in alternative (3), the 6th mass index is not aligned on the amino acid. Then, D[6,6] = D[5,6] - 4 = −43, if −4 is the penalty when a mass index is not aligned. SpecPeptidOMS chooses the alternative with the greatest score that aligns the 6th mass index on amino acid T (score = 7).

several labels of $aa_i$ are found in $Se_c$', one or several mass indexes might be involved in an alignment with two options: (1) the alignment of $aa_i$ continues the alignment of $aa_{i-1}$, leading to an increase in score, and (2) a shift in the spectrum (respectively in the protein) is required to align $aa_i$ after $aa_{i-1}$ (respectively after $aa_{i'}$ with i' + 1 < i), and then, the score is decreased. Between both possibilities, the algorithm selects the best-scored one with a preference for option (1) in the case of a tie. If *D* was effectively a matrix of dimension *n*m*, where *n* represents the number of amino acids of $P^R$, and *m* the number of mass indexes of $Se_c$, *SpecPeptidOMS* would have to fill in 1.15 billion cells to align one spectrum with an estimate of 100 column indexes per $Se_c$' on the human protein database (containing about 11.5 million amino acids). This is not an option as it would take too much time. The central idea behind *SpecPeptidOMS* is to limit the number of cells that can be filled. To achieve this goal, it is important to note that *D* can be seen as a sparse matrix in which most cell values can be inferred from a limited number of cells, referred to as *key-cells*. A cell *D[i,j]* is a key-cell if and only if $aa_i$ is labeled at mass index j. In Figure 3, we colored the key-cells in matrix *D* in blue for the alignment of $Se_c1$' on a fictitious protein. In this example, whereas a complete filling of D would require the calculation of 300 cells (more precisely, 20 × 15, where 20 is the number of mass indexes and 15 is the number of amino acids of the protein), the computation of 28 key-cells only is necessary, if we exclude the initialization of the first row and the first column.

Most of the scores in *D* can be inferred from key-cell scores because when amino acid $aa_i$ is not-found at *j*, the score of *D[i,j]* is equal to the score of the first key-cell *D[i',j]* (possibly itself) met with i'≤ i on mass index *j* minus the *penalty* due to not-found amino acids applied *i - i' - 1* times (Supporting Information Figure S1). Since only key-cell scores need to be computed while *D* is filled, *D* can be compressed as a two-row matrix *D'* storing information on the last key-cell score at each mass index j (*D'[1,j]* contains the score; *D'[0,j]* stores the index of the amino acid i' in $P^R$ where *D'[1,j]* was computed). The evolution of D' along the alignment of Se1 with $P_1^R$, amino acid by amino acid (Supporting Information Table S1).

It is also noteworthy that, similar to *SpecGlobX*, *SpecPeptidOMS* tolerates missing peaks in the alignment due to the


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

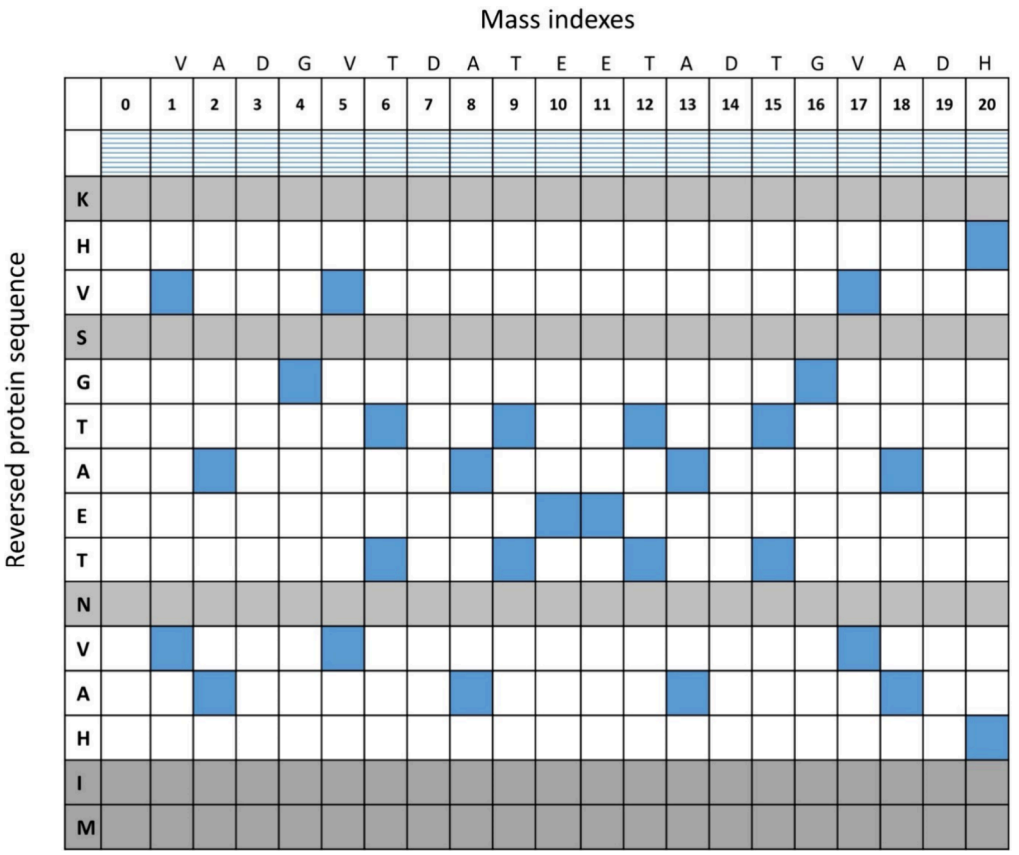



**Figure 3.** Key-cells in the dynamic programming matrix *D*. To align spectrum $Se1_c'$ representing peptide HAVN[0.98]TEATGS[27.99]V (from Figure 1) on the fictitious protein $P_1$, whose sequence is MI**HAVNTEATGSV**HK (the bold part is the peptide represented by *Se1*), *SpecPeptidOMS* only needs to compute the values of the blue-colored key-cells. Because amino acid A in $Se_c1'$ might be aligned on mass indexes 2, 8, 13, and 18 only (see Figure 1d), the row corresponding to A contains 4 key-cells. For readability reasons, at the top of each column, we have added the labeled amino acid corresponding to each mass index. (In this example, a mass index is labeled by only one amino acid, which is not always the case). The gray-colored rows highlight amino acids not-found in $Se_c1'$.

concept of the gap. This concept involves the recognition of amino acid stretches that cannot be recognized individually due to missing peaks, but whose mass corresponds to the mass shift required to align with the next labeled amino acid. The tolerance to missing peaks is given by the parameter *tolPeakMissing*. Technically, when the algorithm computes the score of a key-cell considering the shift with the best score, a gap corresponding to this shift is tested. If the gap is confirmed, it is not scored negatively as a shift should be, but it is scored as an aligned amino acid, i.e., with a bonus of 7 or 10. Gap detections are illustrated later in the article with the spectrum *Se2*.

Due to the transformation of spectra during the preprocessing phase, consisting of adding complementary peaks and potentially leading to double the number of masses in *Se-*, each mass in $Se_c$ has a complementary mass (each mass in black has a complementary mass in black or green in Figure 1b). In other words, each time $aa_i$ is labeled on mass index j in $Se_c'$, it can also be labeled on the mass index of its complementary mass j' except when the lower (respectively upper) bound of the label begins (respectively ends) on the first (respectively last) mass index. However, if $aa_i$ is labeled both on j and j', only one of the two assignments might lead to a correct alignment because in the twin pair (j, j'), only one represents a y-ion that may be aligned on $P^R$ (the other one being the complementary b-ion). Thus, scores are unnecessarily computed for a part of the key-cells in matrix D. To save time, we recommend selecting the *threePeaks* option proposed by *SpecPeptidOMS*. As a result, *SpecPeptidOMS* calculates the key-cell *D[i,j]* only if the label $aa_i$ follows the label $aa_{i-1}$ in $Se_c'$, which means that at least three masses are aligned (i.e., two consecutive amino acids) - or if the label $aa_i$ does not follow any label. The underlying principle of this optimization is that when three masses are aligned, the probability of aligning y-ions and not b-ions is higher (because the amino acid sequence is read from the C-Term to the N-Term). We point out that verifying the *threePeaks* condition is based on an intrinsic spectrum property that can be checked instantly thanks to spectra preprocessing. Then, since the number of computed cells in *D* is nearly divided by two, this option accelerates the spectra processing time, even though the gain is on a much smaller scale than what results from the key-cell concept. The key-cells computed when $Se_c1'$ is aligned on protein $P_1^R$ once the *threePeaks* option is selected are shown in Figure 4. In this example, only 10 keys-cells among 28 are calculated.

As the D' matrices are filled for every pair of protein-spectrum, *SpecPeptidOMS* stores the x best $L_{ss}$SMs—x chosen by the user—corresponding to the x highest scores obtained. An $L_{SS}$SM is defined by the association of an experimental spectrum Se and a subsequence $P^R[i\text{-}\alpha1,j+\alpha2]$, where j is the number of a row containing one of the x highest scores, i is the number of the row where the best-scored alignment begins and $\alpha1$ and $\alpha2$ are the length of small amino acid stretches used to extend the subsequence so that the boundaries of the second


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

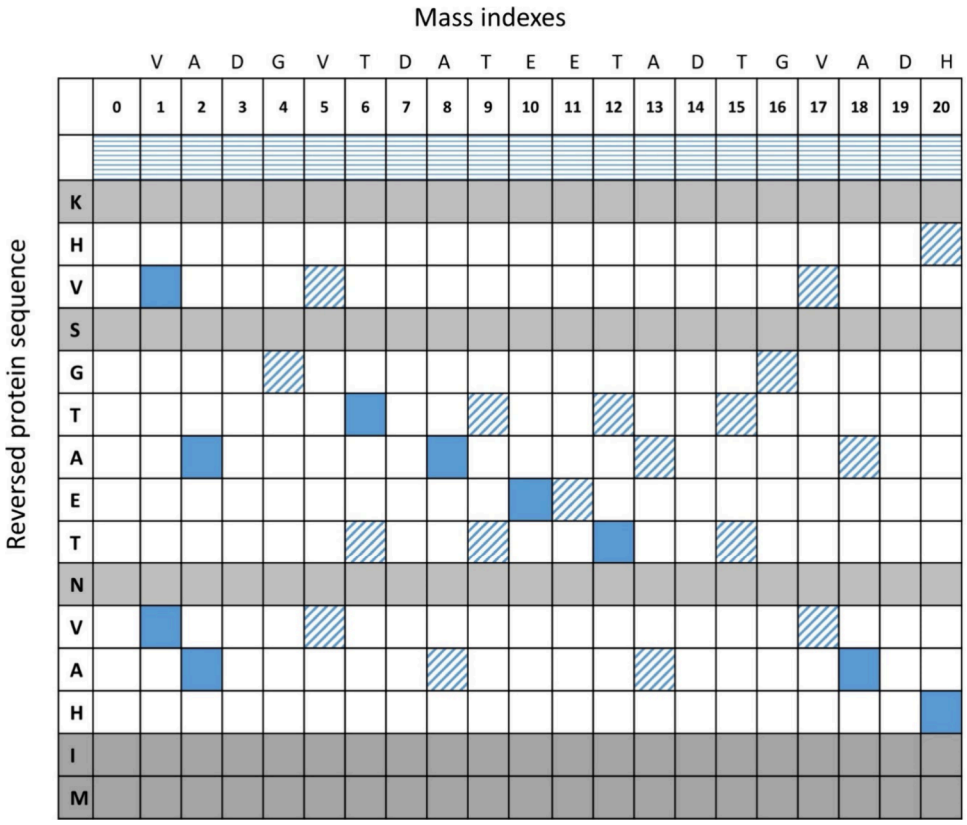



**Figure 4.** Key-cells computed in the dynamic programming matrix $D$ when the *threePeaks* option is set. To align spectrum $Se_c1'$ on a fictitious protein $P_1^R$ MI**HAVNTEATGSV**HK, only the subset of 10 computed cells colored in plain blue is calculated when the *threePeaks* option is set. For instance, even if the two rows corresponding to 'A' contain four key-cells, only two are actually computed. In the first row containing 'A', only key-cells at mass indexes 2 and 8 are computed: at the mass index 8, the label 'A' follows the label 'T' in $Se_c1'$ as it is the case in $P_1^R$, i.e., the lower mass bound of 'A' at mass index 8 is the mass index 6 labeled 'T', so three consecutive peaks are aligned; at mass index 2, the lower mass bound of A was previously deleted because it was not labeled. Then, at mass index 2, 'A' does not "follow" any other amino acid and consequently, this key-cell is computed. On the other side, the key-cells at mass indexes 13 and 18 are not computed because their lower mass bounds are respectively 'E' and 'V', not 'T'. In the second row containing 'A', the key-cell at index 2 is computed for the same reason and the key-cell at index 18 follows the label 'V' as in $P_1^R$, generating the alignment of 3 consecutive peaks.

round alignment can be adjusted. The value of $\alpha 1$ (respectively $\alpha 2$) is computed from the mass shift at the beginning (respectively end) of the alignment if any, divided by the smallest amino acid mass (57.02) and rounded, plus a user-defined parameter (initialized by default to 5). It should be noted that *SpecPeptidOMS* does not require the tracking of paths leading to LssSMs during preliminary alignments; only the index of the row where the alignment begins needs to be retained in memory. This behavior is an additional way to gain speed.

## Algorithm Overview (Part 2): Second Mass Spectra Alignment Round on $L_{ss}$SMs

Because the second alignment round is performed on $L_{ss}$SMs instead of an entire proteome, *SpecPeptidOMS* can spend more time adjusting alignments. First, the *threePeaks* option is no longer considered since it may underestimate the score. Second, the assessment of a gap only for the best shift in the preliminary alignment is extended to all possible shifts. Finally, we added the *PeaksCleaning* option, which can detect some incorrect alignments *a posteriori*. Indeed, when both a peak and its complementary peak are simultaneously aligned on $P^R$, it is highly probable that this alignment is erroneous, as at least one of these peaks corresponds to a b-ion. We therefore decided to implement a correction mechanism capable of detecting and correcting such errors. If the *PeaksCleaning* option is set, two new experimental spectra called *corrected spectra* are built for each pair of complementary peaks by removing one of the pair's peaks from the original spectrum. If several pairs are detected in an alignment, the entire combinatorial of corrected spectra is constructed. Ultimately, the highest-scored alignments of these corrected spectra are retained if their scores reach the minimum threshold required by the user.

## Algorithm Overview (Part 3): Detection of Nonaligned Masses

During the spectrum preprocessing step, we generate complementary peaks based on the relation between a b-ion, its corresponding y-ion, and the mass of the precursor ion that has been fragmented. However, there are situations where the precursor mass referenced in the MS2 spectrum is not the actual mass of the precursor that has been fragmented; this happens, for example, when the charge state of the precursor ion has been wrongly estimated or when the precursor ion has lost a neutral chemical group just before the fragmentation step. In this case, the complementary peaks added during the preprocessing step are shifted from where they should be. More precisely, if the difference between the measured mass and the actual mass is $\Delta_{mass}$, all the complementary peaks undergo a mass shift of $\Delta_{mass}$. To overcome this issue, *SpecPeptidOMS* implements a final step highlighting such $\Delta_{mass}$ shifts and corrects the alignments accordingly.


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

**Table 1. *SpecPeptidOMS* Parameters with Default Values**

| Parameter Name | Value | Comment |
|---|---|---|
| CModif | 57.02146 | Fixed modification on cysteines |
| nbMinPeaks | 30 | Only spectra above that number of peaks are processed |
| nbMinAAFound | 20 | Only spectra above that number of labels are processed |
| nbSelectedPeaks | 60 | Number of the most intense peaks filtered per spectrum |
| accuracy | 0.02 | Fragment accuracy |
| threePeaks | true | Option (true/false) to accelerate the preliminary alignment |
| nbLssSMSaved | 5 | Number of Located subsequence Spectrum Matches ($L_{ss}SM$) per spectrum |
| nbResultsReturned | 1 | Number of interpretations per spectrum returned to the user |
| peaksCleaning | true | Option (true/false) to correct specific alignments in the second alignment round |
| tolPeakMissing | 4 | Maximum number of missing peaks accepted to score a gap as an alignment (in both alignment rounds) |
| minScenarioScore | 30 | Only interpretations with a score over this threshold are returned to the user |
| nbResultsAtOnce | 500 | Number of spectra alignments written at a time |
| surplus | 5 | Number of amino acids that extend the $L_{ss}SM$ for the second alignment |
| tolPeakMissingFirstCol | 5 | Maximum number of amino acids that can be jumped without penalty at the beginning of $L_{ss}SM$ |
| filterLssSMOnScore | 0.9 | Excludes $L_{ss}SM$ with a score below the best score multiplied by this factor when several are returned per spectrum by the preliminary alignment |
| nbThreads | 10 | Number of threads |
| certainlyFound | 10 | Elementary scoring in dynamic programming when a labeled amino acid is aligned without any mass shift, and both y-ion and b-ion are present |
| found | 7 | Elementary scoring in dynamic programming when a labeled amino acid is aligned without any mass shift and only y-ion or b-ion is present |
| certainlyFoundWithShift | −6 | Elementary scoring in dynamic programming for a shift, and both y-ion and b-ion are present |
| found WithShift | −8 | Elementary scoring in dynamic programming for a shift, and only y-ion or b-ion is present |
| notFound | −4 | Elementary scoring in dynamic programming when the amino acid is not aligned (not-found or an alternative is preferred) |

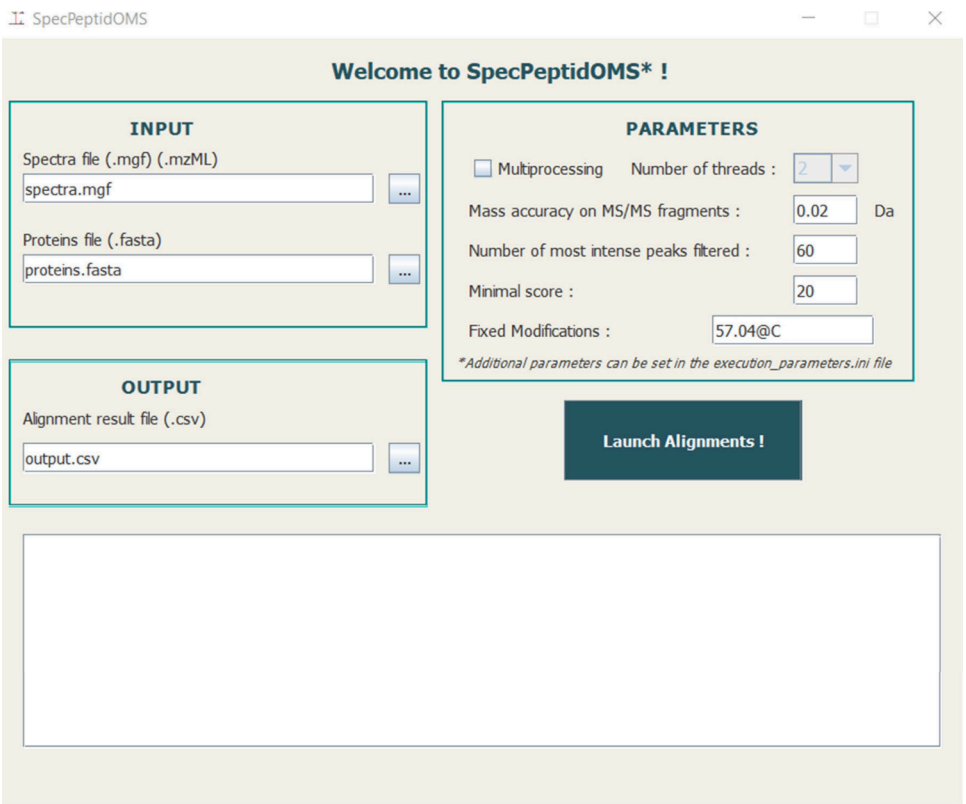



**Figure 5.** *SpecPeptidOMS* graphical user interface.

We hypothesized that each shift in the alignment could reveal the presence of one $\Delta_{mass}$. Consequently, *SpecPeptidOMS* assesses the alignments for all of the theoretical spectra generated by sequentially adjusting the mass of the precursor by each of the shifts. Among the set of $\Delta_{mass}$ values that are tested, the one with the best alignment score is selected (in the case of a tie, the one that relies on the maximum number of peaks is chosen).

The shifts at the extremities of the LssSM are specific. Indeed, when a shift $s$ is located at the beginning (respectively at the end) of an alignment, its mass might not correspond exactly to $\Delta_{mass}$ but to $\Delta_{mass}$ minus the mass of several amino


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

Table 2. Comparison of the Number of Identifications for the *Synthetic_HLA2* Spectra Data Set between *MaxQuant* and *SpecPeptidOMS*[a]

| | | Both search engines | | | | |
|---|---|---|---|---|---|---|
| MaxQuant only | SpecPeptidOMS only | #ident | #included | #diff | Total interp. | Exec time (in min) |
| 8,341 (16.5%) | 10,309 (20.4%) | 17,818 (35.3%) | 3,090 (6%) | 10,980 (21.8%) | 50,338 | 239 |

[a]Columns 1 and 2 display how many spectra are interpreted by only one of the two search engines. Columns 3 to 5 indicate how many sequences identified by *SpecPeptidOMS* are identical (#ident), different (#diff), or represent a sub-sequence (#included) from the *MaxQuant* sequences when both search engines identify spectra. Column 6 is the total number of interpreted spectra (sum of columns 1 to 5). The last column gives the execution time for *SpecPeptidOMS*. All percentages were calculated in relation to the total number of interpretations.

acids $aa_i...aa_j$ preceding (respectively following) the aligned sequence in $P^R$. As the number of amino acids to be considered is unknown, all possible amino acids stretch $aa_i...aa_j$ so that $\Delta_{mass} = s$ - mass $(aa_i..aa_j)$ is positive or equal to −1 Da are tested to determine which $\Delta_{mass}$ generates the best alignment score. However, a $\Delta_{mass}$ = s - mass $(aa_i..aa_j)$ equal to 1 or −1 Da is preferred, even if the related alignment does not return the best score, because isotopic mass errors are frequent. At the end, the alignment corresponding to the selected $\Delta_{mass}$ is compared to the original alignment (the one without $\Delta_{mass}$) and if it is strictly better, the value of $\Delta_{mass}$ and its corresponding alignment are returned to the user in addition to the original alignment. More precisely, a better alignment means that it reaches a higher score and relies on more shared peaks (the number of peaks in common between the experimental spectrum and its interpretation as a peptide) than does the original alignment does.

### SpecPeptidOMS Parameters

The *SpecPeptidOMS* parameters used for the analyzed data sets are summarized in Table 1.

### Execution Time Measures

SpecPeptidOMS was executed on a laptop with an Intel i7−8850H CPU (2.6 GHz) and 16 GB of memory allocated to the Java Virtual Machine, running under Windows 10.

We run MaxQuant v2.6.2.0 using the standard protocol described in Tyanova et al.[24]

## RESULTS

### SpecPeptidOMS Usage

We implemented *SpecPeptidOMS* as a standalone Java application that can be run through a graphical user interface (GUI) displayed in Figure 5, or in command-line mode. The GUI includes all main options required to run the alignments and a log area to report processing information. As input, *SpecPeptidOMS* takes a list of MS2 spectra representing the experimentally fragmented peptides (tryptic or nontryptic) and a list of protein sequences, possibly as large as a whole proteome. *SpecPeptidOMS* reads spectra files in MGF or mzML format (parsed by the JMZReader library[25]), and the list of protein sequences in the fasta format.

*SpecPeptidOMS* is an open-source project publicly available on GitHub (https://github.com/bibs-lab/SpecPeptidOMS).

For each alignment, *SpecPeptidOMS* returns the following information: (1) the interpreted peptide, a string describing the alignment with a syntax similar to what is used in *SpecGlobX*, (2) the names of at most three proteins containing the peptide (possibly followed by the string ". . ." indicating a larger list), (3) the score, and (4) the number of shared peaks between the filtered spectrum and the peptide. Selecting the *nonAlignedMass* option may result in additional information containing the improved alignment and the nonaligned mass value if any.

### SpecPeptidOMS Speed and Memory Performances

It took about 4 h (239 min) to run *SpecPeptidOMS* on the *Synthetic_HLA2* data set with the standard configuration (Table 1). To evaluate the major parameters influencing execution speed, we conducted a series of measurements on a subset of 5,000 spectra extracted from *Synthetic_HLA2*. Apart from the parameters that may filter many spectra (*nbMinPeaks* or *nbMinAAFound*), only the parameters that affect the first alignment computation time significantly impact the total execution time. Indeed, the total execution time depends mainly on the number of computed key-cells, which is divided by 10,000 between the first and the second alignment including the postprocessing (less than 10 million for the second alignment versus more than 100 billion for the first alignment). Consequently, three parameters act on the execution time: *threePeaks*, *accuracy* and *nbSelectedPeaks* (Supporting Information, Tables S2, S3, and S4). Importantly, except for these three parameters, all parameters can be changed without degrading the performance.

Memory usage is not a limiting factor with *SpecPeptidOMS*. Most of the data structures are allocated once per thread and related to the size of the largest protein and to the maximum number of filtered peaks per spectrum. However, we recommend avoiding an accumulation of results in memory by setting the *nbResultsAtOnce* parameter to a relatively low value (a backup of every 500 spectra by default).

### SpecPeptidOMS Accuracy on the Synthetic_HLA2 Data Set

We benchmarked the accuracy of *SpecPeptidOMS* interpretations on the *Synthetic_HLA2* spectra data set (Supporting Information, Table S5). *SpecPeptidOMS* loaded 57,221 spectra with at least 30 peaks and 20 labeled amino acids representing the fragmentation of 941 nontryptic synthesized peptides, which mainly end with a glutamine and range from 7 to 17 amino acids (approximately 14 on average). The best advantages of *Synthetic_HLA2* are that synthesized peptides have known sequences, and spectra interpretations using *MaxQuant* are downloadable. To challenge *SpecPeptidOMS*, we merged the synthetic peptide database with the human proteome database.

We compared the results obtained by *SpecPeptidOMS* with those obtained by *MaxQuant*, a well-established search engine in the community. We point out that the presented results in Table 2 are not directly comparable, because *MaxQuant* was run against a limited database (i.e., the synthetic peptide database) with only methionine oxidation and protein N-terminal acetylation considered as possible modifications. In doing so, the results using *MaxQuant* benefit from the user's


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

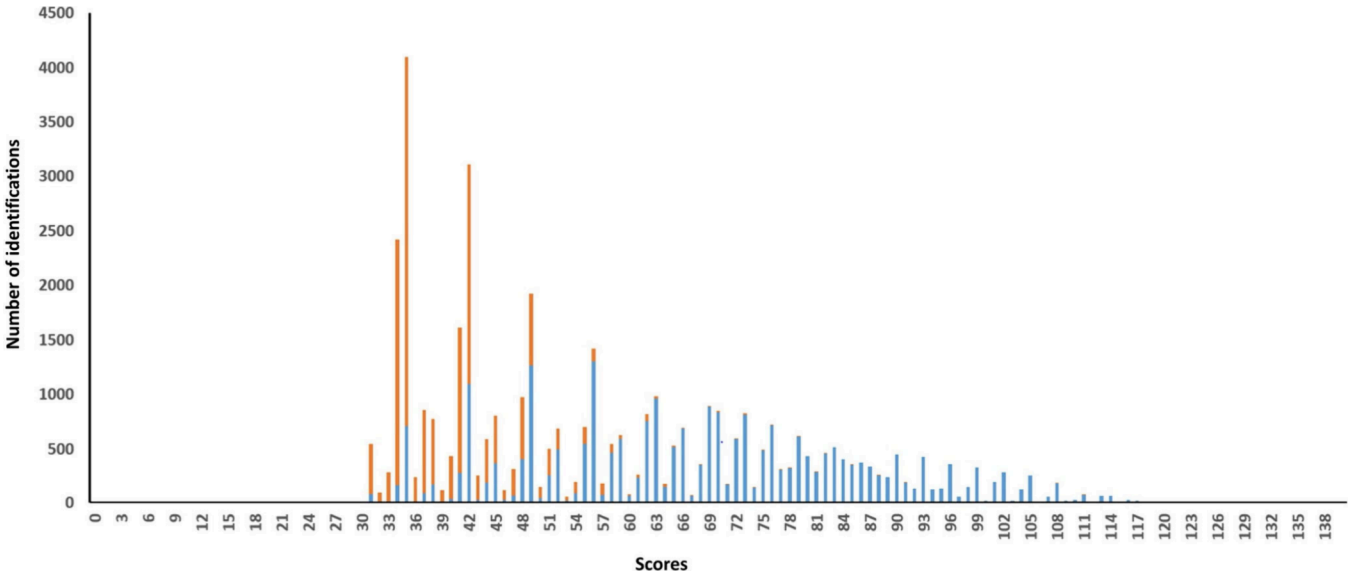



**Figure 6.** The ratio between true and false interpretations of the Synthetic_HLA2 data set according to*SpecPeptidOMS* alignment scores. False interpretations are in orange, whereas true interpretations are in blue.

in-depth knowledge of what makes up the synthetic sample, so we estimated those interpretations as more reliable. In contrast, we challenged *SpecPeptidOMS* using the synthetic database mixed with the whole human proteome under real-life consistent conditions of a lesser-known sample analysis. The raw juxtaposition of the results is not fair for *SpecPeptidOMS*, but if we bear this in mind when comparing the results, it provides significant insight into the performances of *SpecPeptidOMS*.

For information, executing *MaxQuant* on the same database as *SpecPeptidOMS* took 21 h (1260 min) on a laptop using 10 threads and returned 34,286 identifications at FDR of 1% (instead of 40,331 when using the restricted database). This was expected as *MaxQuant* has to model a multitude of theoretical spectra resulting from all of the peptides along the proteins, ranging from the minimum to the maximum length considered.

About 12% of the MS2 spectra were not identified by either of the two search engines, probably due to their poor quality. The largest proportion of the remaining spectra is consistently interpreted by both search engines, with 41% of spectra leading to identical or included interpretations. The proportion of discrepancies concerns around 22% of spectra, and 37% of the spectra are interpreted by only one search engine. Two other parameter sets were tested and compared to *MaxQuant* (Supporting Information, Table S6).

We investigated the factors contributing to the discrepancies between the two search engines using the standard setting. To ensure a meaningful comparison, we need to ascertain the score threshold at which *SpecPeptidOMS* provides interpretations upon which we can rely upon. Given that synthetic peptide sequences account for only 0.3% of the total merged protein sequence volume, a spectrum is unlikely interpreted as a synthetic peptide by chance, at least if the peptide sequence is long enough (above six amino acids). Thus, we considered that a sequence from the synthetic peptide database is likely true, and that a sequence from the human protein database is likely false. *SpecPeptidOMS* returns 41,997 spectra interpretations, of which 24,709 are true (i.e., representing synthetic peptide sequences of at least 6 amino acids), i.e., 59% of the results. The distribution between true and false interpretations of the scores (Figure 6) highlights a rapid change in the ratio around a score value of 50. Indeed, 90% of the identifications are true when the score is higher than 50 (42% of the spectra), and 97% when it is higher than or equal to 60 (31% of the spectra).

Because it is difficult to find the origins of discrepancies when both search engines return a low score, we concentrate our efforts on the set of "well-interpreted" spectra, i.e., spectra that obtain a score > = 90 using *MaxQuant* (22,366 spectra) or a score ≥60 with *SpecPeptidOMS* (18,556 spectra, 54% of the interpreted spectra). The discrepancies between both search engines interpreting those spectra were classified into three categories according to their scores: (1) the first category includes spectra obtaining a high identification score with each search engine, but returning a different interpretation; (2) the second category (about 11% of of *Synthetic_HLA2*) merge spectra with no *SpecPeptidOMS* interpretation (2,620 spectra) or with a score lower than 60 (3,602 spectra) while simultaneously having a high *MaxQuant* score (≥90); (3) last, the third category (about 5% of *Synthetic_HLA2*) includes spectra with a *SpecPeptidOMS* reliable score (≥60) while *MaxQuant* returns a low score (<90, 592 spectra), or no interpretation (2,328 spectra).

The first category is small (294 spectra, 0.4% of *Synthetic_HLA2*), since when both search engines reach a high score (13,341 spectra, 23% of *Synthetic_HLA2*), the interpretation is almost always consistent, thereby reinforcing our confidence in *SpecPeptidOMS* (remind that *SpecPeptidOMS* runs using a more challenging search space). As high scores are synonymous with good-quality interpretations for both search engines, we manually investigated what could explain the differences in interpretation. First, *Synthetic_HLA2* contains chimeric spectra: these spectra can be considered artifacts in MS2 analysis because they combine the fragments of two (or more) distinct peptides coisolated during the fragmentation process. As each of the search engines interpreted one of the two coisolated peptides, both can be considered to have interpreted the spectra correctly. Scan 46656 illustrates a chimeric spectrum interpreted differently (Figure S2). Second,


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

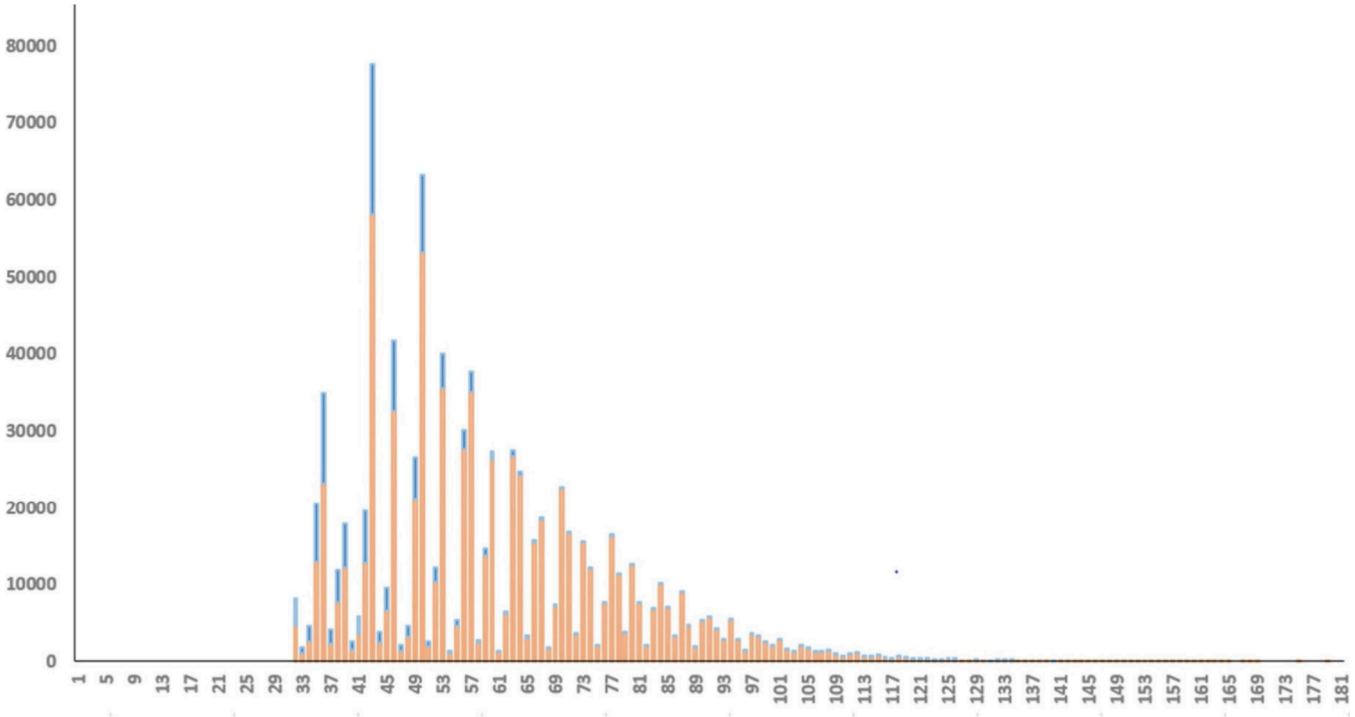



**Figure 7.** Distribution of tryptic vs nontryptic peptides identified by *SpecPeptidOMS* on the HEK293 data set (without specifying the digestion enzyme). The orange bar shows the number of identified tryptic peptides, while the blue bar shows the number of nontryptic peptides.

the presence of close sequence peptides complicates the question of whether an interpretation is incorrect. For instance, scan 27787 is interpreted as EDLRSWTAAD**M**[**15.99**]AAQ by *MaxQuant*, and as EDLRSWTAAD[**T**][**45.99**]AAQ by *SpecPeptidOMS* since both peptides EDLRSWTAAD**M**AAQ and EDLRSWTAAD**T**AAQ are in the synthetic database. These interpretations are distinct, although both alignments are identical. In *SpecPeptidOMS*, because no assumption is made about modifications, oxidized methionines are considered and penalized as any other modification, unlike in *MaxQuant*. Given the information provided by the mass spectrum alone, we could consider that both interpretations are correct; however, if we add information about sample preparation, the interpretation returned by *MaxQuant* would be the most probable. Third, we must admit that *SpecPeptidOMS* has an intrinsic weakness toward a particular type of peptide we call a "low-complexity peptide". Those peptides are composed of repeated sequences or long stretches of repeated amino acids that generate a bias toward spectra that share this stretch. Scan 13721 is a perfect illustration of this situation: it is interpreted as AGAAAAAAHTAIQQ by *MaxQuant* with strong confidence, and as [274.13]-AAAAA[67.04]AAAAAA[110.05] by *SpecPeptidOMS* (Supporting Information, Figure S3). Because the peaks were filtered, *SpecPeptidOMS* did not recognize the C-terminal end of the peptide. Afterward, the central offset of 67.05 Da allows a series of 'A' to be aligned on y-ions, followed by a series of 'A' aligned inadvertently on b-ions, resulting in an incorrect identification. Approximately 25% of the discrepancies in the first category are due to this phenomenon. However, low-complexity peptide identification is easily discernible in the results. Users are strongly advised to exercise caution when validating these low-complexity peptides, particularly if the alignment displays a central shift. Fortunately, such interpretations usually represent a small proportion of the results. We also point out that this issue reflects a sort of amino acid mirroring effect between b-ions and y-ions in the original spectrum and is not a consequence of the addition of complementary peaks (these latter peaks are corrected by the *peakCleaning* option). The difficulty generated by low-complexity peptides is shared by all OMS methods relying on alignments, including manual annotation of spectra.

In the second category, a low *SpecPeptidOMS* score results from the difficulty of aligning one (or several) contiguous series of amino acids possibly interrupted by one (or several) shifts corresponding to a protein subsequence in the database. *SpecPeptidOMS* tolerates missing peaks with the gap approach but only to a certain extent. By contrast, *MaxQuant* uses the precursor mass to select peptide candidates, making it more tolerant in the absence of peaks, but preconceived ideas may sometimes lead to unfortunate identifications.

The third category demonstrates one of the main advantages of *SpecPeptidOMS*, namely, its ability to identify modifications and/or neutral losses in spectra without preconceived ideas. Although *Synthetic_HLA2* has been prepared to avoid peptide modifications, *SpecPeptidOMS* reveals an unambiguous labile modification of 56.06 Da on a series of spectra (Supporting Information, Figure S4). Moreover, peptide synthesis can be imperfect. For instance, *SpecPeptidOMS* returns the alignment PQDLAAA[T][-101,05]AKLVGQ for scan 37008, which suggests that the threonine amino acid is missing (Supporting Information, Figure S5). Additionally, many *MaxQuant* uninterpreted spectra exhibit an ion parent mass error greater than tolerable in parameters. Finally, for a significant part of the spectra, we did not elucidate why *MaxQuant* missed some interpretations that seem unambiguous (examples in the Supporting Information, Figure S6).

In conclusion, except for a weakness linked to the OMS approach concerning a small, limited subset of easily identifiable interpretations, identifications obtained by *Spec-*


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

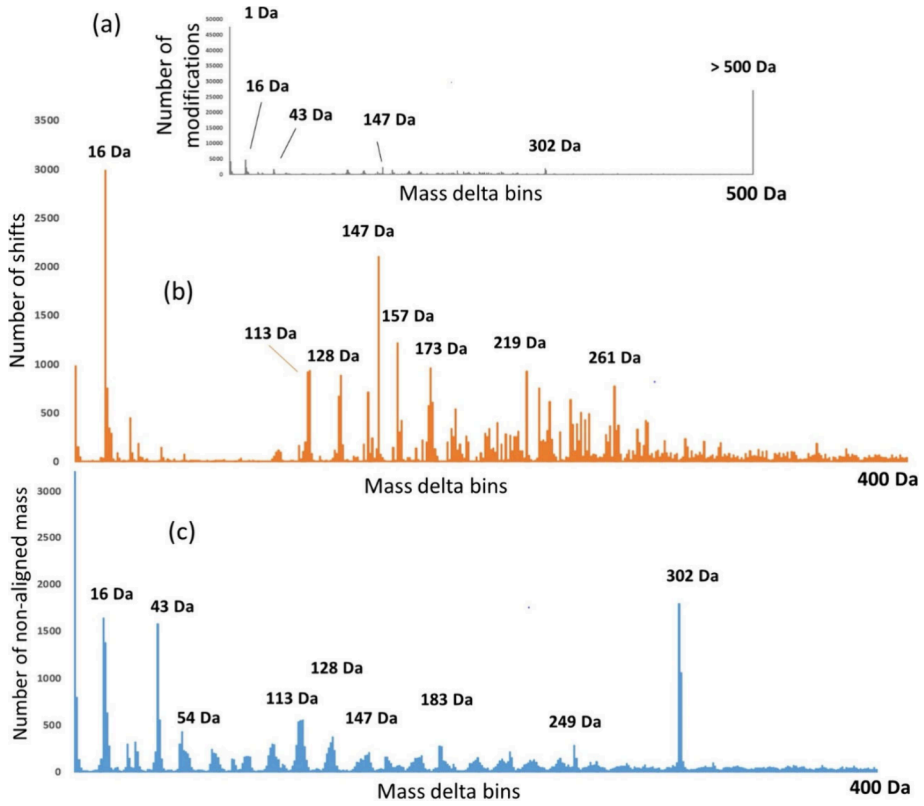



**Figure 8.** Distribution of mass shifts and nonaligned masses returned by *SpecPeptidOMS* when interpreting the HEK293 data set. We selected only interpretations with a score ≥60 to compute the distribution. The number of modifications (respectively shifts or nonaligned masses) are counted in bins centered on integer values + - 0.5 Da. (a) the general histogram overview shows the major modification masses. (b) the detailed histogram in orange represents the distribution of mass shifts between 0 and 400 Da. To observe minor peaks, the peak at one Dalton is removed (15,854 shifts); (c) the detailed histogram in blue displays the number of nonaligned mass shifts (the peak at one Dalton is removed).

*PeptidOMS* are consistent with those returned by *MaxQuant*. This is a promising result since *SpecPeptidOMS* was run in a drastically larger search space: larger database, no a priori modifications, including neutral losses. *SpecPeptidOMS* overcomes the difficulty of unknown peptide boundaries in terms of reasonable computation times and benefits from all of the advantages of an OMS approach. Consequently, *SpecPeptidOMS* can interpret spectra carrying unanticipated modifications in peptidomics.

## Alignments Highlighted in the HEK293 Data Set

As a second evaluation, we ran *SpecPeptidOMS* on a proteomics data set generated from HEK293 cells downloaded from PRIDE (PXD001468, 24 files) without specifying the tryptic digestion. This data set contains a larger diversity of spectra than *Synthetic_HLA2* and has been used several times to evaluate open modification search methods. The processing lasted about 4 h per file containing an average of 45,000 spectra on our laptop. For an initial effectiveness estimate, we count the number of identified peptides ending by K or R according to the score (Figure 7). Although identification of nontryptic peptides is expected (the last peptide of proteins does not always end with K or R) and a significant proportion of semitryptic peptides is to be anticipated, we expect a low proportion of nontryptic peptides. This is observed above the score of 50, suggesting that most of the spectra are completely interpreted and correctly located on the human proteome without using digestion information.

Next, we focused on the modifications suggested, whether in the form of shifts or nonaligned masses. We selected identifications obtaining a score of at least 60 (316,274 spectra) and clustered the mass shifts (respectively the nonaligned masses) into 1 Da wide mass bins between 0 and 500 Da, except the masses above 500 Da merged in the last bin.

The distribution of mass shifts and nonaligned masses is displayed in Figure 8. Overall, the most frequent modifications follows those previously documented in Chick et al.[20]

Interestingly, modifications such as oxidation (16 Da) are identified either as a shift or as a nonaligned mass. Actually, if a spectrum includes the superposed signature of the fragmented peptide with and without the modification, *SpecPeptidOMS* favors labile modification interpretation, avoiding the shift penalty. Besides, among the most frequent shifts, we recognize the mass of several amino acids (113 Da; 128 Da). Those shifts are located mostly at the N-Term of the peptide and likely reveal the presence of genetic variants not represented in the human database. For instance, scan 43199 is interpreted as [114.04]IEINPDSAQPYK. If 'N' had preceded IEINPDSAQPYK, *SpecPeptidOMS* would have returned the complete sequence. Actually, the preceding amino acid is 'A', suggesting a mutation. On the other hand, the frequent shift at 147 Da is associated with peptides beginning with an oxidized methionine.

Finally, to assess the effectiveness of the postprocessing step, we focused on the interpretation of a subset of 77 spectra from the *b1922_293T_proteinID_02A_QE3_122212* file, identifying the peptide *pep* = DATNVGDEGGFAPNIIENK, as we had done in Prunier et al.[14] Running successively *SpecOMS* to


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

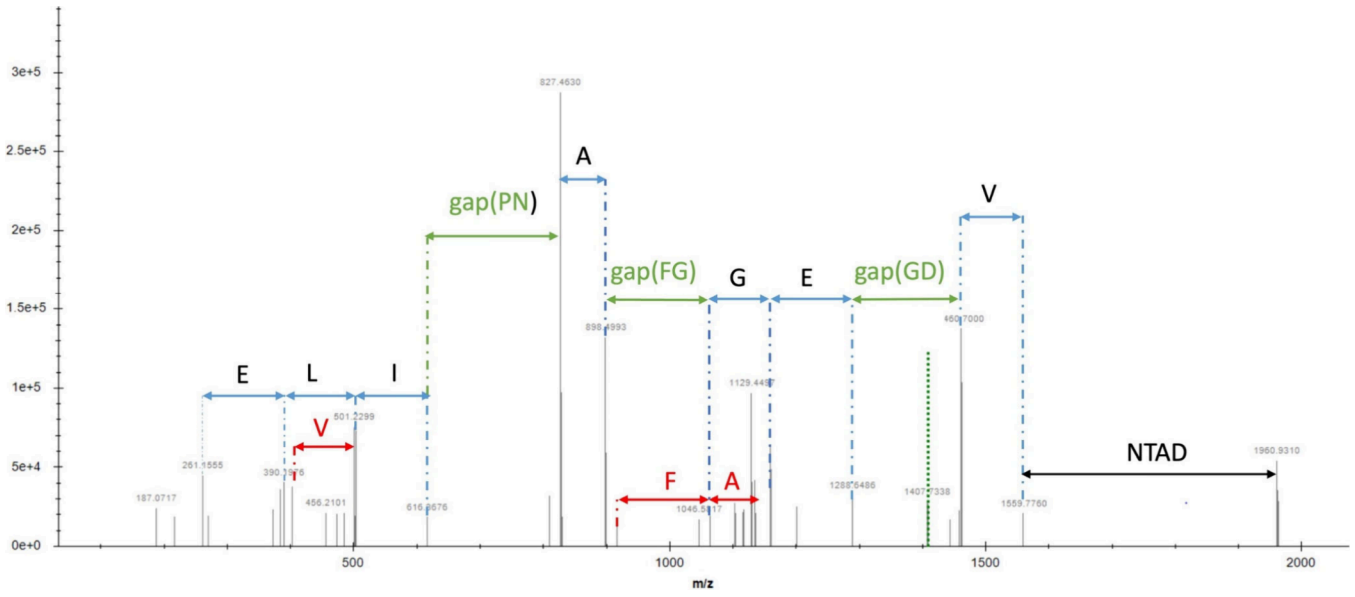



**Figure 9.** Annotated Spectrum *Se2* with labeled amino acids and gaps to show which masses are used to align *Se2* on pep = DATNVGDEGGFAPNILENK. Labeled amino acids based on y-ions are in blue, labeled amino acids based on b-ions are in red, and amino acid stretches aligned as gaps are displayed in green. The stretch in black (NTAD) is interpreted by the post-treatment step.

obtain PSMs, and *SpecGlobX* to improve alignments had previously revealed a wide variety of modifications spread over 44 masses between −500 and 2065 Da, with numerous nonaligned masses. Moreover, the comparative interpretation with *MODPlus*[26] and *MSFragger*[27] is available as Additional File 2 in Prunier et al.[14] (This spectra subset is included with the *SpecPeptidOMS* executable for demonstration). 72 out of 77 spectra are identified as *pep*, while the 5 spectra remaining interpretations are low-scored (scores ≤51). Although this data set is particularly complex, *SpecPeptidOMS* can reveal the same shifts and nonaligned masses as did the association of *SpecOMS* and *SpecGlobX* working on tryptic peptides.

To illustrate some of the *SpecPeptidOMS* strengths, we detail the alignment of scan 54025 on peptide *pep* (Figure 9). First, to challenge *SpecPeptidOMS*, we complicate the identification process by filtering only the 50 most intense peaks to increase the number of missing peaks (otherwise, we used the standard parameters). Additionally, since *Se2* has a nonaligned mass, amino acids labeled from complementary b-ions are useless in this particular case. Finally, only 7 amino acids are labeled on Se2. However, thanks to the concept of gap, *SpecPeptidOMS* aligns *Se2* on *pep* obtaining the alignment string [2664.05]-V[G][D]EG[G][F]A[P][N]ILE[N][K] whose score is 49 (7 amino acids ILE, A, EG and V are aligned with y-ions peaks, giving a score of 7*7 = 49 because a gap is scored to 0). Notably, without this concept of gap, each stretch of at least 2 not-found amino acids would generate a shift and, thus, a too-low score. Afterward, during the postprocessing, *SpecPeptidOMS* extends the sequence with "DATN," increasing the score and the number of shared peaks, leading to a nonaligned mass of 2262.9 (the fragmentation between G and F is highlighted by the nonaligned mass), which gives the alignment string [D][A][T][N]V[G][D]EGGFA[P][N]ILE-[N][K]. Then, *Se2* likely results from a dimerization combined with a neutral loss of 301.99 Da (Supporting Information, Figure S7).

## CONCLUSION

Our objective when we designed *SpecPeptidOMS* was to develop a method that would align peptide fragmentation spectra directly onto whole, undigested protein sequences. In contrast to pure *de novo* methods, our idea was to use protein sequences as "frames", to restrict the search space; but unlike conventional database search engines, not to generate theoretical spectra.

We have successfully addressed the challenge associated with the dynamic programming approach on which the algorithm is based, namely, an excessive execution time. The results obtained on two different data sets (one containing nontryptic peptides whose interpretation is approximately known, and the other corresponding to a large-scale proteomic data set already analyzed by several well-established software) are convincing of the ability of *SpecPeptidOMS* to interpret spectra corresponding to peptides whose extremities are unknown, and possibly carrying modifications. Thus, *SpecPeptidOMS* appears as a promising algorithm to interpret spectra in peptidomics, a field in which satisfactory tools are still lacking. In addition, *SpecPeptidOMS* allows an open modification search in peptidomics without impacting performance.

Although the most appropriate application of *SpecPeptidOMS* is peptidomics, the software can also be applied to classical proteomics data sets (i.e., tryptic peptides). In such a context, *SpecPeptidOMS* is unlikely to compete with conventional or OMS tools for the identification of tryptic or semitryptic peptides, or PTMs on these peptides (both in terms of computational time and in terms of number of identifications). Nevertheless, we believe that SpecPeptidOMS can be useful for interpreting MS2 spectra that remain unidentified because all potential interpretations without any form of filtering are tested, with the particularity, compared to other algorithms, including OMS, of not cleaving the protein sequence into peptides.

We encourage the community to test *SpecPeptidOMS*: it is an easy-to-use software with few parameters to set up that does not require installation. With its command-line mode,


https://doi.org/10.1021/acs.jproteome.4c00870
J. Proteome Res. 2025, 24, 2159−2172

*SpecPeptidOMS* is easy to integrate into a workflow. For large-scale studies, users should consider additional tools to obtain robust statistics to validate the results. Similarly, various tools (not yet available) would be invaluable to help the user interpret shifts and nonaligned mass as chemical modifications.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The data of this study is available in the ProteomeTools project with accession numbers PXD021013 (files 03035a_GA3-TUM_HLA2_3_01_01-3xHCD-1h-R1.raw, msms_db.csv and peptides.txt) and PXD001468 (24 spectra data sets generated from HEK293 cells). The human database used in all the tests was downloaded from UniProt (UP000005640_9606.fasta). *SpecPeptidOMS* is an open-source project and is publicly available on GitHub https://github.com/bibs-lab/SpecPeptidOMS.

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.4c00870.

Extended view of the programming matrix D when the spectrum Se1 is preliminary aligned on the fictitious protein $P_1$ (Figure S1); Evolution of matrix D' (compressed version of matrix D) throughout the alignment (Table S1); *SpecPeptidOMS* execution time according to the *threePeaks* parameter (Table S2); *SpecPeptidOMS* execution time according to the *accuracy* parameter (Table S3); *SpecPeptidOMS* execution time according to the *nbSelectedPeaks* parameter (Table S4); Comparison of the number of identifications for Synthetic_HLA2 spectra data set between *MaxQuant* and *SpecPeptidOMS* according to two different parameter settings (Table S6); Interpretations of scan 46656 (*Synthetic_HLA2* data set) as a chimeric peptide (Figure S2); Interpretation of scan 13721 (*Synthetic_HLA2* data set) (Figure S3); *SpecPeptidOMS* interpretation of scan 57856 (Figure S4); *SpecPeptidOMS* interpretation of scan 37008 (*Synthetic_HLA2* data set) (Figure S5); *SpecPeptidOMS* interpretation of scan 44853 (Figure S6); *SpecPeptidOMS* interpretation of scan 54025 (Se2) as DATNVGDEGGFAPNILENK with an nonaligned mass (Figure S7) (PDF)

Table S5: Synthetic_HLA2 interpretation using *SpecPeptidOMS* (XLSX)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Géraldine Jean** − *Nantes Université, CNRS, LS2N, UMR 6004, F-44000 Nantes, France;* orcid.org/0000-0002-1534-2682; Email: Geraldine.Jean@univ-nantes.fr

### Authors

**Émile Benoist** − *Nantes Université, CNRS, LS2N, UMR 6004, F-44000 Nantes, France*

**Hélène Rogniaux** − *INRAE, PROBE Research Infrastructure, BIBS Facility, F-44300 Nantes, France; INRAE, UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France;* orcid.org/0000-0001-6083-2034

**Guillaume Fertin** − *Nantes Université, CNRS, LS2N, UMR 6004, F-44000 Nantes, France;* orcid.org/0000-0002-8251-2012

**Dominique Tessier** − *INRAE, PROBE Research Infrastructure, BIBS Facility, F-44300 Nantes, France; INRAE, UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.4c00870


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the French National Agency for Research, ANR, through the ANR project *PeptidOMS* (ANR-24-CE45-3296).

## ■ REFERENCES


(1) Ahrné, E.; Müller, M.; Lisacek, F. Unrestricted identification of modified proteins using MS/MS. *Proteomics* **2010**, *10* (4), 671−686.

(2) Ramazi, S.; Zahiri, J. Posttranslational modifications in proteins: resources, tools and prediction methods. *Database (Oxford)* **2021**, *2021*. DOI: 10.1093/database/baab012.

(3) Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaíno, J. A. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **2016**, *13* (8), 651−656.

(4) Bogdanow, B.; Zauber, H.; Selbach, M. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. *Mol. Cell Proteomics* **2016**, *15* (8), 2791−2801.

(5) Hellinger, R.; Sigurdsson, A.; Wu, W.; Romanova, E. V.; Li, L.; Sweedler, J. V.; Süssmuth, R. D.; Gruber, C. W. Peptidomics. *Nature Reviews Methods Primers* **2023**, *3* (1), 1−21.

(6) Schrader, M. Origins, Technological Advancement, and Applications of Peptidomics. *Methods Mol. Biol.* **2024**, *2758*, 3−47.

(7) Fan, K. T.; Hsu, C. W.; Chen, Y. R. Mass spectrometry in the discovery of peptides involved in intercellular communication: From targeted to untargeted peptidomics approaches. *Mass Spectrom Rev.* **2023**, *42* (6), 2404−2425.

(8) Schulz-Knappe, P.; Schrader, M.; Zucht, H. D. The peptidomics concept. *Comb Chem. High Throughput Screen* **2005**, *8* (8), 697−704.

(9) Foreman, R. E.; George, A. L.; Reimann, F.; Gribble, F. M.; Kay, R. G. Peptidomics: A Review of Clinical Applications and Methodologies. *J. Proteome Res.* **2021**, *20* (8), 3782−3797.

(10) Martini, S.; Solieri, L.; Tagliazucchi, D. Peptidomics: new trends in food science. *Current Opinion in Food Science* **2021**, *39*, 51−59.

(11) Menschaert, G.; Vandekerckhove, T. T.; Baggerman, G.; Schoofs, L.; Luyten, W.; Van Criekinge, W. Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J. Proteome Res.* **2010**, *9* (5), 2051−2061.

(12) Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell Proteomics* **2012**, *11* (4), M111.010587.

(13) De La Toba, E. A.; Anapindi, K. D. B.; Sweedler, J. V. Assessment and Comparison of Database Search Engines for Peptidomic Applications. *J. Proteome Res.* **2023**, *22* (10), 3123−3134.

(14) Prunier, G.; Cherkaoui, M.; Lysiak, A.; Langella, O.; Blein-Nicolas, M.; Lollier, V.; Benoist, E.; Jean, G.; Fertin, G.; Rogniaux, H.; Tessier, D. Fast alignment of mass spectra in large proteomics datasets, capturing dissimilarities arising from multiple complex modifications of peptides. *BMC Bioinformatics* **2023**, *24* (1), 421.

(15) David, M.; Fertin, G.; Rogniaux, H.; Tessier, D. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *J. Proteome Res.* **2017**, *16* (8), 3030−3038.

(16) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10* (4), 1794−1805.
(17) Perez-Riverol, Y.; Bai, J.; Bandla, C.; Garcia-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaino, J. A. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research* **2022**, *50* (D1), D543.
(18) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; Yu, P.; Schlegl, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **2017**, *14* (3), 259−262.
(19) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918−920.
(20) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743−749.
(21) Wilhelm, M.; Zolg, D. P.; Graber, M.; Gessulat, S.; Schmidt, T.; Schnatbaum, K.; Schwencke-Westphal, C.; Seifert, P.; de Andrade Kraetzig, N.; Zerweck, J.; Knaute, T.; Braeunlein, E.; Samaras, P.; Lautenbacher, L.; Klaeger, S.; Wenschuh, H.; Rad, R.; Delanghe, B.; Huhmer, A.; Carr, S. A.; Clauser, K. R.; Krackhardt, A. M.; Reimer, U.; Kuster, B. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* **2021**, *12* (1), 3346.
(22) Uniprot Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.
(23) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. Statistical Characterization of Ion Trap Tandem Mass Spectra from Doubly Charged Tryptic Peptides. *Anal. Chem.* **2003**, *75* (5), 1155−1163.
(24) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11* (12), 2301−2319.
(25) Griss, J.; Reisinger, F.; Hermjakob, H.; Vizcaíno, J. A. jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics* **2012**, *12* (6), 795−798.
(26) Na, S.; Kim, J.; Paek, E. MODplus: Robust and Unrestrictive Identification of Post-Translational Modifications Using Mass Spectrometry. *Anal. Chem.* **2019**, *91* (17), 11324−11333.
(27) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14* (5), 513−520.


https://doi.org/10.1021/acs.jproteome.4c00870
*J. Proteome Res.* 2025, 24, 2159−2172