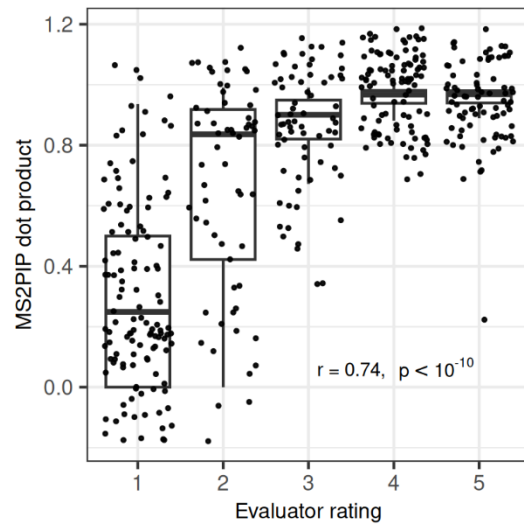
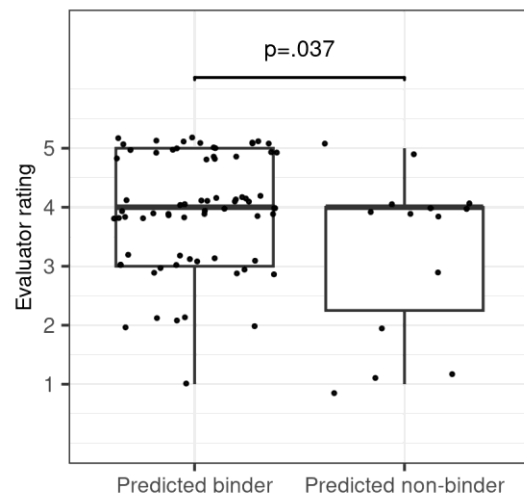


## Supplementary Information

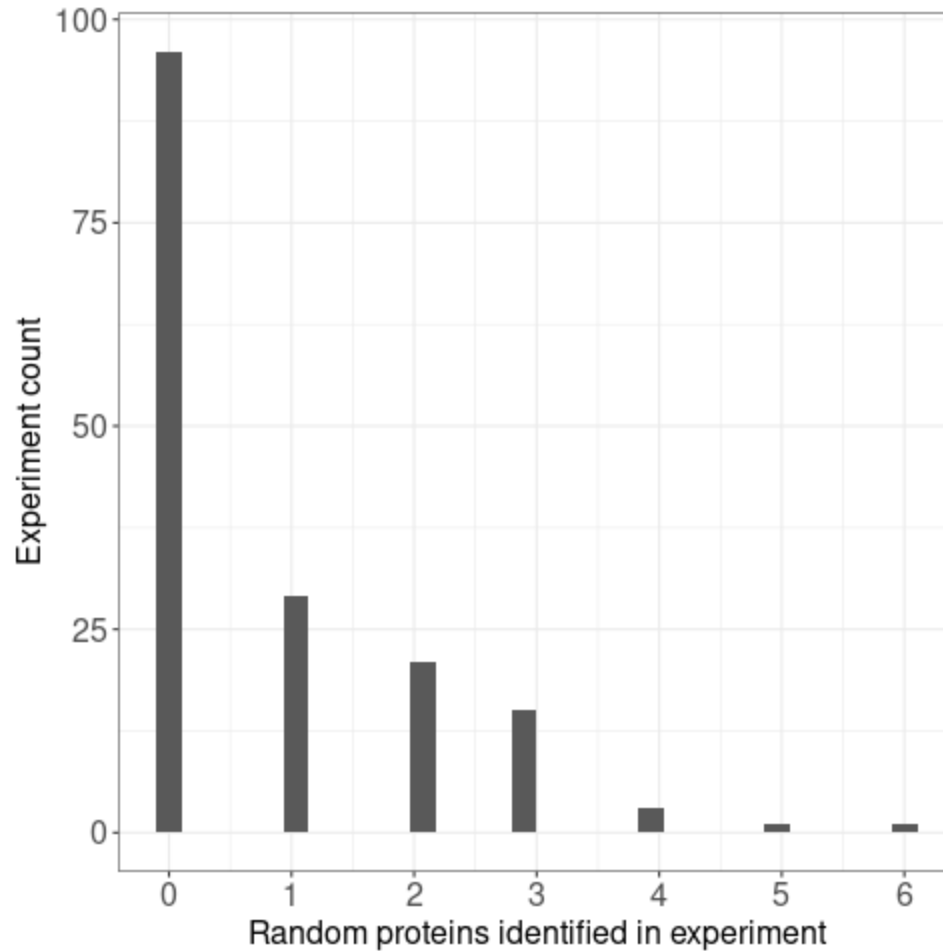
### Supplementary Figures



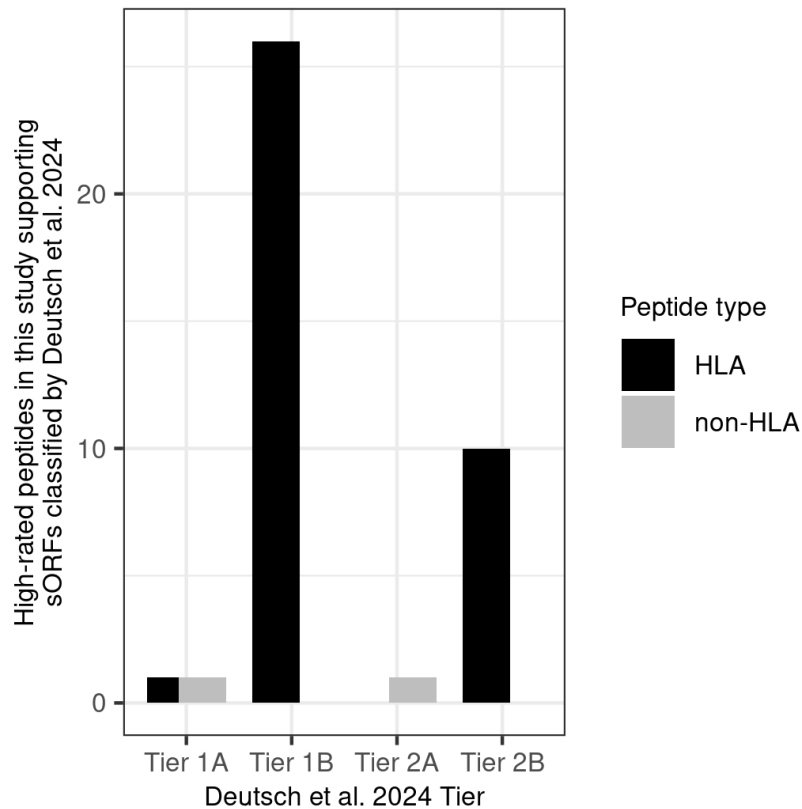
**Supplementary Figure 1: Evaluator rating is strongly correlated with dot product between observed spectra and spectra predicted by MS2PIP.** MS2PIP was used to generate predicted spectra for each evaluated PSM ( $n = 428$  PSMs). The dot product between the predicted and observed spectra is shown for each PSM, with PSMs grouped by manual evaluator rating. The correlation between dot products and ratings is given.



**Supplementary Figure 2: Peptides that are predicted to bind HLAs are more highly rated by evaluators.** For each evaluated immunopeptidomics peptide from Ouspenskaia et al. 2021, Martinez et al. 2020, or Chong et al. 2020, HLA binding was predicted using NetMHC. Any peptide meeting the NetMHC criteria for weak or strong binder to an HLA allele present in the cell type was considered a predicted binder. The distribution of evaluator ratings among predicted binders and peptides not predicted to bind HLAs is shown ( $n = 85$  PSMs associated immunopeptidomics peptides). Difference in group means is tested using a two-sided permutation test.



**Supplementary Figure 3: Proteomic searches for random proteins in human MS datasets falsely report detections when canonical proteins are excluded from the protein database.** Histogram showing the number of random proteins detected among studies when MSGF+ was used to detect a sample of 10 randomly constructed proteins against a human MS dataset with 166 experiments. Proteins were considered detected if they had a peptide with a reported q-value <1%. This plot demonstrates that, in the absence of genuine detection of any protein in the database, it is common for a few proteins to be reported with q-value <1%. This is because the q-value for a given PSM is estimated as the number of decoys with confidence score above that of the PSM divided by the number of targets with confidence scores above the PSM. Under the null hypothesis of zero genuine detections, it is equally likely that a target or decoy has the highest confidence score; when it is a target it will be assigned a q-value of 0. For instance, Chothani et al.<sup>4</sup> employed a two-stage strategy to detect sORF products. In the first stage, the UniProt human proteome was used as the sequence database. For each MS experiment, any spectra that matched with a peptide at the 1% FDR threshold was removed from the spectra file. In the second stage, the sORF list was used as the sequence database against the modified spectra file, and any sORF product with a peptide identified at the 1% FDR threshold was considered to be detected. Since all annotated proteins were removed from the database in the second stage, and there may be no unannotated proteins detectable in the sample, the conditions of no genuine protein detections are potentially met. As shown by this plot, under these conditions false positives are expected.



**Supplementary Figure 4: Peptides highly rated in this study that also supported ORFs classified in Deutsch et al. 2024.** The number of high-rated peptides (ratings of 4 or 5) analyzed in this study that were also validated in Deutsch et al. 2024 and used as support for ORFs of various tiers classified in that paper. Peptides are divided by whether they were found in HLA immunopeptidomics experiments or non-HLA conventional proteomics experiments. The ORF tier definitions, taken from Deutsch et al. 2024, are as follows. Tier 1A: ORF supported by at least two non-nested peptides from conventional proteomics experiments as well as Ribo-Seq data. Tier 1B: ORF supported by at least two non-nested peptides from HLA immunopeptidomics experiments as well as Ribo-Seq data. Tier 2A: ORF supported by at least one peptide from conventional proteomics experiments as well as Ribo-Seq data. Tier 2B: ORF supported by at least one peptide from HLA immunopeptidomics experiments as well as Ribo-Seq data.

## Supplementary Tables

Supplementary Table 1: Properties of all PSMs reported to support unannotated protein detections that were considered in this study.

PMID	Citation	Year	Source	Search engine used in study
35393574	Cao	2022	Spreadsheet provided by authors	MaxQuant
35788065	Bogaert	2022	Table S3, "With confidence" tab	Mascot
35841888	Chothani	2022	Table S10, "MS hits" tab	MS-GF+
36171426	Duffy	2022	Table S2 for the ORF list. Authors could not provide spectra identifiers. Spectra identifiers came from replication attempt using the parameter files ending mqpar.xml at the PRIDE repository PXD035950.	MaxQuant
34276900	Cai	2021	Was not able to get data from authors	Proteome Discoverer
34193551	Douka	2021	Data provided by authors	Comet
33510483	Prensner	2021	Table S14	Spectrum Mill
34663921	Ouspenskaia	2021	Tables S3, S6, S8, S9, S12 for PSMs, but these do not include modifications. Modification information was taken from data export files in the MassIVE database with identifier MSV000084787.	Spectrum Mill
32139545	Chen	2020	The list of detected ORFs is from Figure S4A. The PSMs were then taken from the output file at PXD014031: txt_FullProteome_sORF.rar by matching to that list.	MaxQuant
32157095	Chong	2020	Table S3	Comet and MaxQuant
31819274	Martinez	2020	Spreadsheets provided by authors	pFind 3 Open-pFind and DTASelect
31155234	van Heesch	2019	Data provided by authors	MaxQuant
31340039	Lu	2019	Table S6	Mascot, MaxQuant and X!Tandem

## Appendix 1: PSM rating scheme and examples

The following scheme was given to evaluators as a basis for rating each PSM. Below, example PSMs are given together with an explanation for their rating provided by an evaluator. Each PSM can be visualized on ProteomeCentral (<https://proteomecentral.proteomexchange.org/usi/>) using the USI.

5 - Excellent. A very good match that shows peaks for nearly every residue except perhaps b1 (and thus the order of the two utmost N-terminal ions is unclear) and is not contaminated by a cofragmented precursor. Really solid evidence for the peptide. Example:  
mzspec:PXD021482:20200724\_cell\_10:scan:6083:APQSPGPAPPPASSGR/2

4 - Good enough even for an extraordinary detection. Perhaps one or two peaks missing, perhaps some contamination, but seems like a good match. Example:  
mzspec:PXD014058:20181120\_HCT116\_P-ACN\_up\_14:scan:35747:GGQSLPTTMWSPVK/2

3 - Not good enough. It might be right. But it might well be something else fairly close. Incomplete coverage of the residues. This PSM would be fine if it were a common albumin peptide, but the bar for a non-canonical ORF is higher. Examples:

mzspec:PXD020079:20180504\_QEh1\_LC1\_QC\_JMI\_HLAIp\_HROG17\_2\_R2:scan:6918:AVAGSRGD  
KSLR/3

mzspec:PXD010154:01698\_A02\_P018021\_S00\_N09\_R1:scan:11204:ASEIQSTGGQRDPQPER/3

mzspec:PXD004894:20141216\_QEp7\_MiBa\_SA\_HLA-I-p\_MMf\_6\_2:scan:29039:KPRLPIYGL/2

mzspec:MSV000080527:M20151203\_HLA\_A2402\_75millionceq\_biorep1\_techrep1:scan:46144:YS  
LSLQILF/2

2 - Wrong with a high quality spectrum. Clearly not the correct interpretation, although perhaps close, but this spectrum probably has a good alternative explanation. Usually, some high unexplained peaks near but not exactly at where there ought to be a peak if the interpretation were correct. Example:

mzspec:PXD004894:20141215\_QEp7\_MiBa\_SA\_HLA-I-p\_MMf\_15\_1:scan:34668:MPRMALVYHTA/3

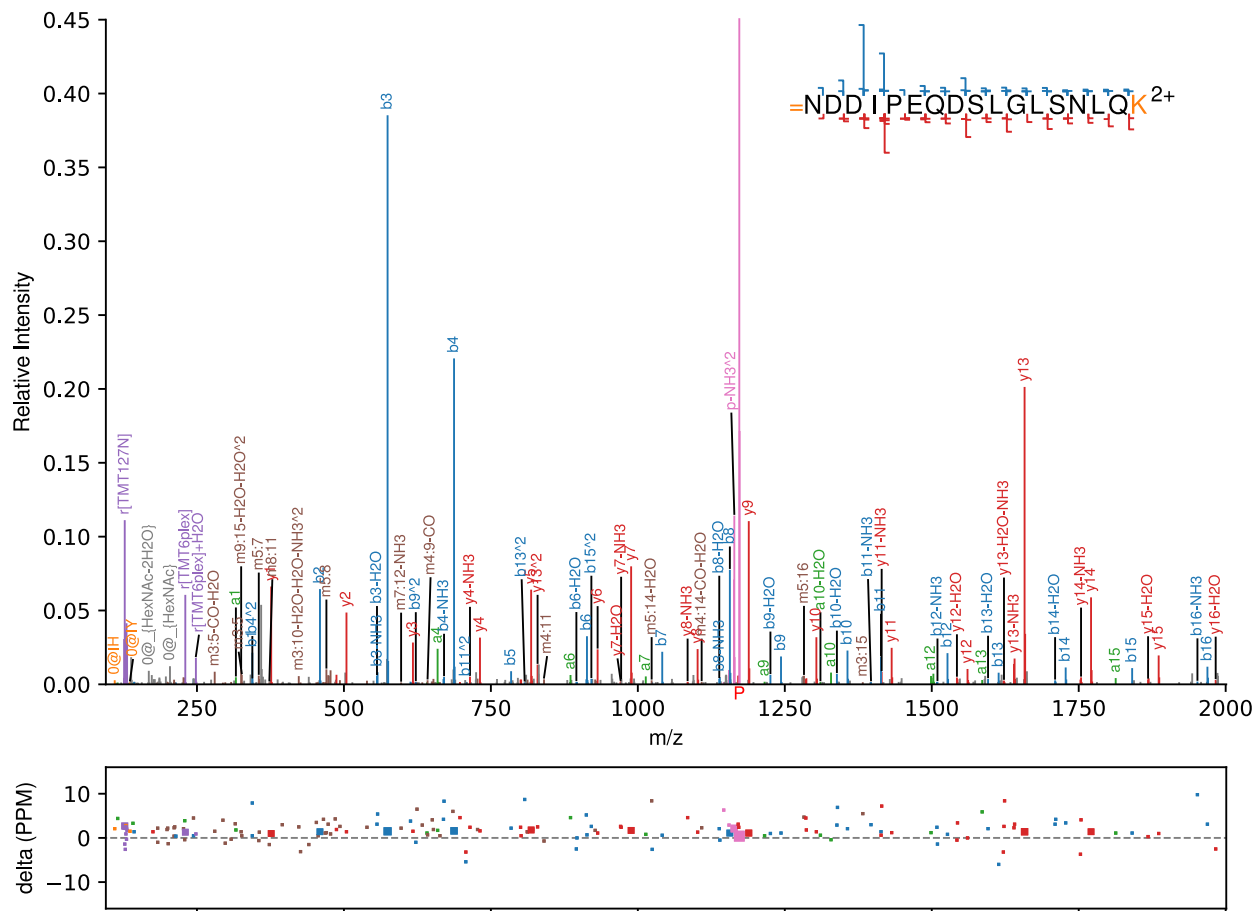
1 - Poor quality spectrum. Not enough information to be sure about any id. Or perhaps a clearly blended spectrum where it's hard to be sure of any id. Example:

mzspec:PXD019643:170421\_AM\_AUT01-DN14\_BoneMarrow\_W6-32\_10\_\_DDA\_2\_400-  
650mz\_msms23\_standard:scan:26200:SVWLSPPPA/2

### Example PSMs with consistent results from the evaluators

Example of a 5 star rating: Both reviewers gave it a 5.

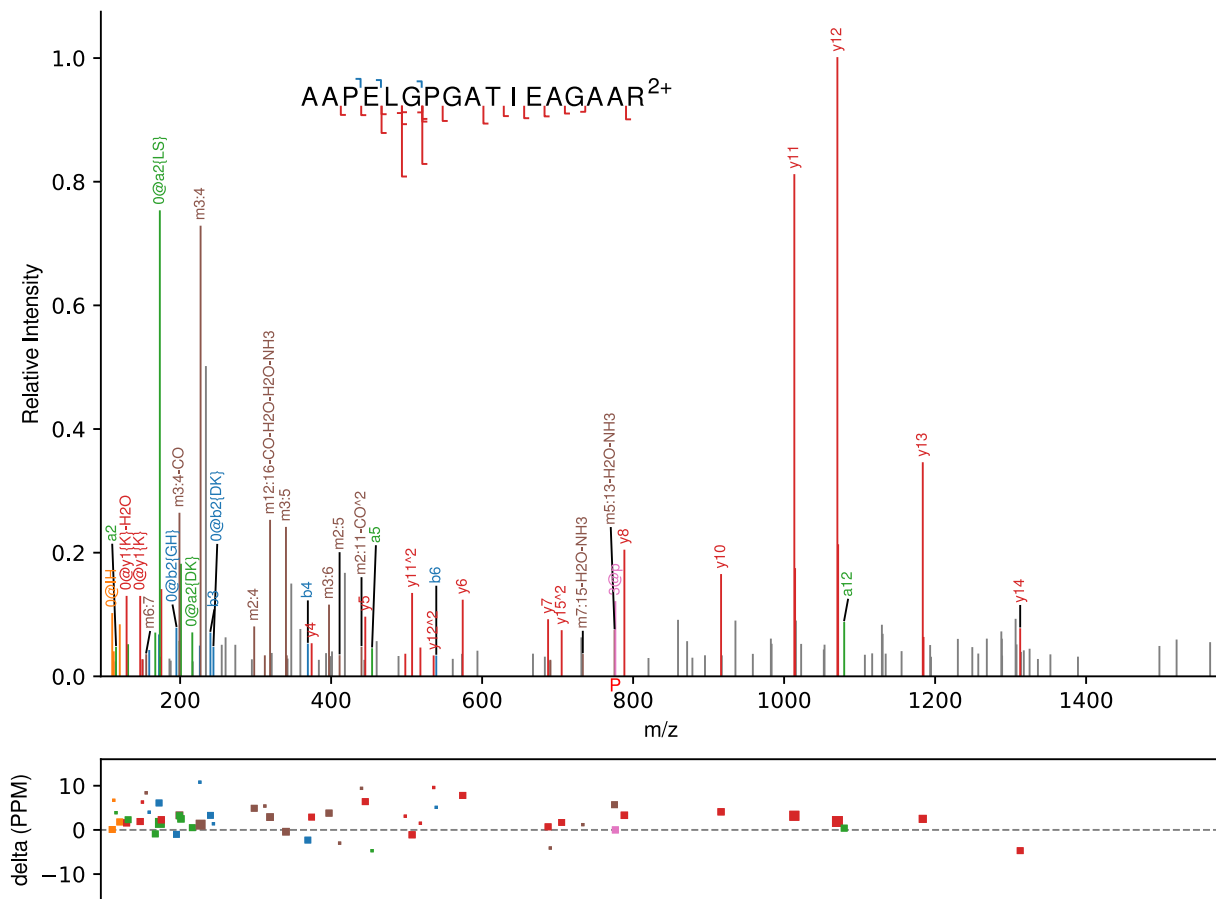
mzspec:PXD999953:09CPTAC\_UCEC\_W\_PNNL\_20180222\_B3S1\_f03:scan:33089:[TMT6plex]-  
NDDIPEQDSLGLSNLQK[TMT6plex]/2



This PSM presents a very strong match with full coverage from both ends. This 64 amino acid protein was highlighted in Kim et al. 2014.<sup>74</sup> It is currently in UniProtKB/TrEMBL as Q9HB66 but not in UniProtKB/Swiss-Prot.

Example of a 4 star rating: Both reviewers gave it a 4.

mzspec:PXD026880:VOT16-2132:scan:3865:AAPELGPGATIEAGAAR/2

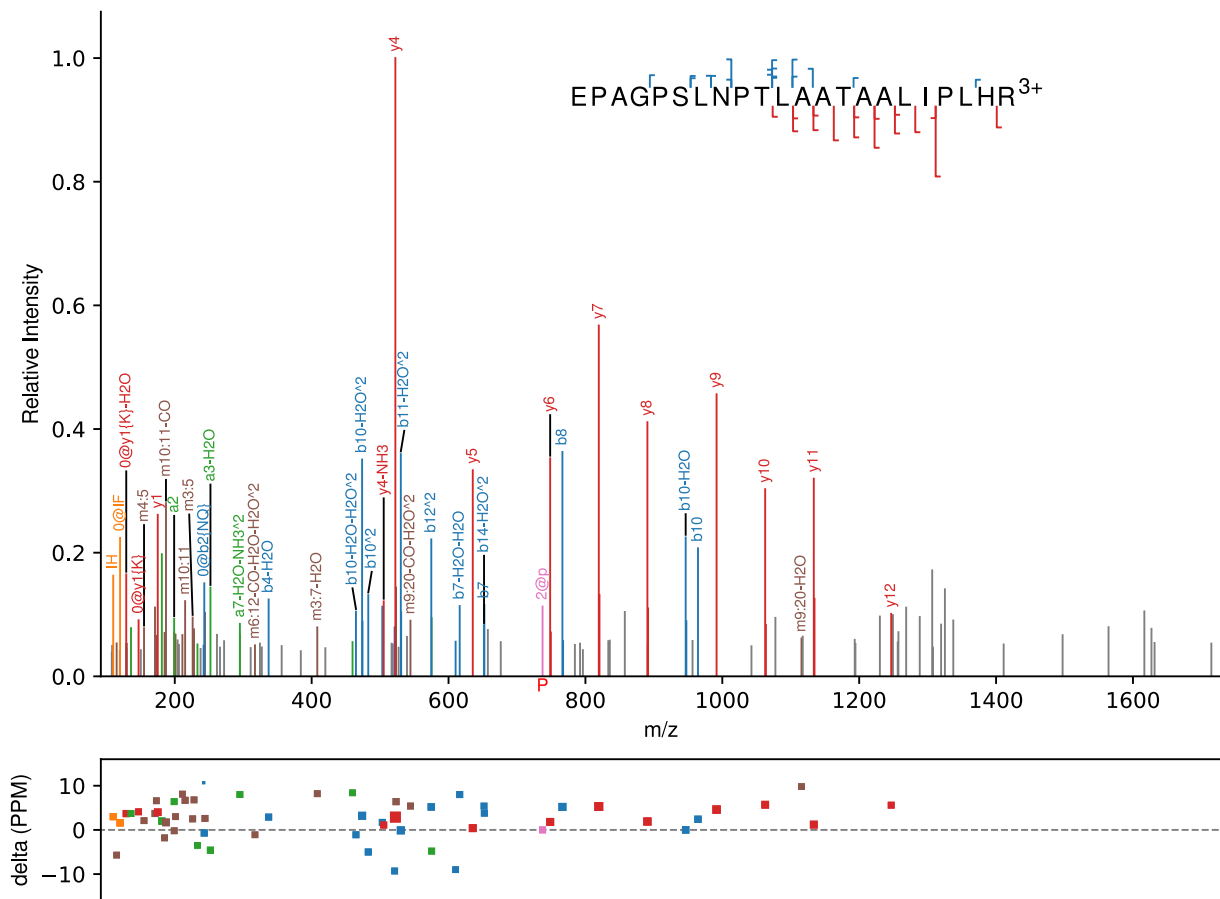


This PSM provides nearly complete coverage in y ions, although there are some gaps. The b ions are very weak, but that is not surprising given the sequence. Signal to noise (as estimated by the ratio of the tallest to smallest peak) is decent, but weaker than the PSMs rated 5. The precursor m/z value is exactly as expected.

There are a few major peaks that are not easily explained except by a y-ion series of a contaminating peptide ending in LSK, explaining peaks at 147.1311, 235.1455, and 347.2301. This reduces confidence slightly. For these reasons, this PSM does not rate a 5, but is good evidence for the peptide.

Example of a 3 star rating: All 3 reviewers gave it a 3.

mzspec:PXD026880:VOT16-2132:scan:9332:EPAGPSLNPTLAATAALIPLHR/3



There is good evidence for the second half of the peptide, but evidence is almost entirely lacking for the first half of the peptide. The peptide sequence is likely at least partially correct but may have a different N terminus. If this were an annotated protein, it would be sufficient, but not for a claim of a novel protein.

Example of a 2 star rating: Both reviewers gave it a 2.

mzspec:PXD014031:20190104\_QX4\_AnBr\_SA\_IPSC\_Peptidome\_Fraction\_12:scan:94981:SLEGLIPSSSVVGK/  
2

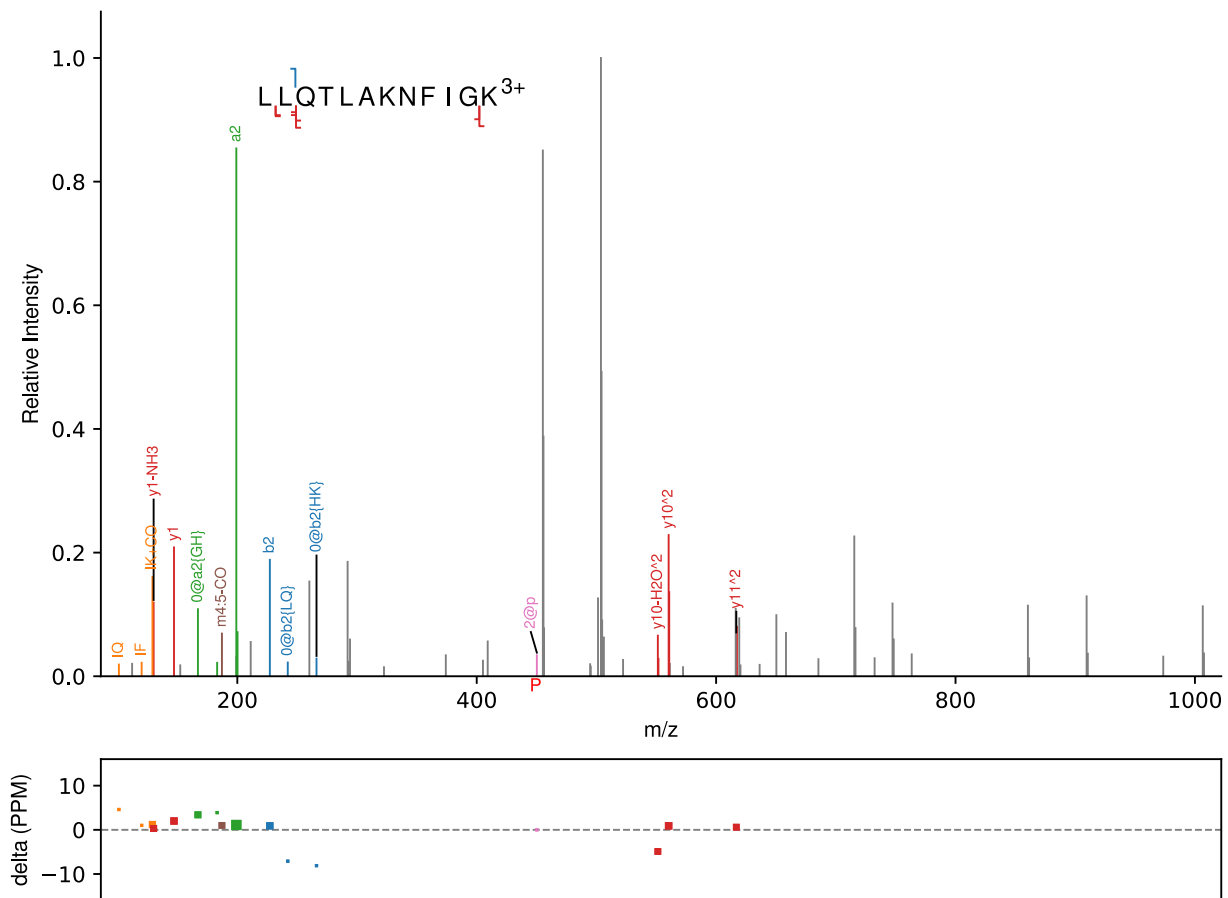






Example of a 1 star rating: Both reviewers gave this spectrum a rating of 1.

mzspec:PXD006675:20160721\_QEp2\_SoDo\_SA\_LC12-13\_RV8-frac2:scan:65697:LLQTLAKNFIGK/3



This is a low signal-to-noise spectrum that is very poor evidence for the claimed peptide and may be too low quality to confidently identify any peptide.

Manual interpretation of the spectrum reveals a peptide that fits far better:

mzspec:PXD006675:20160721\_QEp2\_SoDo\_SA\_LC12-13\_RV8-frac2:scan:65697:LLLPHGVDQLLK/3

explaining most peaks in the spectrum, but it is unclear where that peptide might derive from, as it does not map to known proteins, and there are still gaps in coverage.

## Appendix 2: Mass Spectrometry Microprotein Detection Guidelines

### Mass Spectrometry Microprotein Detection Guidelines

Version 0.2.0 – June 17, 2025

General guidelines for all manuscripts:		
√	Loc	1. Complete this MS Microprotein Detection Guidelines checklist
Data deposition and reference proteome guidelines		
		2a. Deposit all MS proteomics data to a ProteomeXchange repository as a “complete” submission.
		2b. Include analysis reference files (search database, spectral library, transition list, etc.) in submission.
		2c. Provide the PXD identifier(s) in the manuscript abstract.
		2d. Provide the reviewer login credentials if the dataset is not yet public.
		3. Use the most recent version of the HPP Target List plus UniProtKB reviewed isoforms plus contaminants as the reference canonical proteome for all informatics analyses.
FDR-related guidelines		
		4a. Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels.
		4b. Report the PSM-, peptide-, and protein-level FDR values along with the total number of expected false positives at each level, using precision appropriate to the uncertainty in computed FDR.
		4c. Present large-scale results thresholded at equal to or lower than 1% protein level global FDR, but not more than 10% local FDR.
		4d. If any large-scale datasets are individually thresholded and then combined, calculate the new, higher peptide- and protein-level FDRs for the combined result.
Guidelines for claims of new protein detections		
		5a. If using DDA mass spectrometry for such claims, present high mass-accuracy, high-SNR, and clearly annotated spectra to ensure PSMs are correct. Scrutinize spectra for missing and extra peaks.
		5b. Present high mass-accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the spectra supporting the claims. Compute normalized backbone ion dot products between target spectra and synthetic peptide spectra, as well as between target spectra and predicted spectra.
		5c. Provide Universal Spectrum Identifiers (USIs) for all natural, synthetic, and predicted peptide-spectrum matches that support such claims, ideally as a supplementary data table.
		6. If using SRM verification for such claims, present target traces alongside synthetic heavy-labeled peptide traces, demonstrating co-elution and closely matching fragment mass intensity patterns.

		7. If using DIA MS, then, if the data are analyzed with XICs, apply the above SRM guidelines (6); if the data are analyzed by extracting deconvoluted spectra, apply the above DDA guidelines 5a-5c.
		8. Even when very high confidence peptide identifications are demonstrated, consider alternate mappings of each peptide to canonical proteins. Consider isobaric sequence/mass modification variants, all known SAAVs, and unreported SAAVs.
		9. Support Tier 1 (1A digest, 1B HLA) claims via two or more distinct uniquely-mapping, non-nested peptide sequences of length $\geq 9$ amino acids with the above evidence. When 2 peptides overlap, the total extent must be $\geq 18$ amino acids. 100% coverage, except for an initiating methionine, may be considered Tier 1 evidence. Tier 2A (digest) or Tier 2B (HLA) evidence may constitute just a single uniquely-mapping peptide of length $\geq 9$ amino acids. Tier 3 evidence (not meeting Tier 2A or 2B evidence) may be discussed but is generally not considered sufficient for a claim of detection.

See extended description for each of the above items on pages 2 - 4 below.

## Extended Detail on Checklist items:

The following pages provide some additional detail on the intentionally terse one-page checklist. Users should read these extended descriptions before using the checklist.

### 1. Complete this MS Data Protein Detection Guidelines checklist.

#### 2. Guidelines for data repository deposition

- a. **Deposit all MS proteomics data to a ProteomeXchange repository as a complete submission.** All depositions should be “Complete” submissions instead of “Partial” submissions. ProteomeXchange deposition should be completed prior to submission of the manuscript to the journal. Synthetic peptide MS runs should also be deposited and clearly marked as such.
- b. **Include analysis reference files (search database, spectral library, transition list, etc.) in submission.** Include all supplemental data files used in the analysis. Include software parameter files if relevant.
- c. **Provide the PXD identifier(s) in the manuscript abstract.**
- d. **Provide the reviewer login credentials if the dataset is not yet public.** Reviewer login information at the repository should be provided in the manuscript if the dataset is not already publicly released.

### 3. Use the most recent version of the HPP Target List plus UniProtKB reviewed isoforms plus contaminants as the reference canonical proteome for all informatics analyses.

Informatics analysis should always be presented in comparison with the most recent canonical proteome references, rather than older versions thereof. The most recent versions of the HPP Target List, UniProtKB isoforms, and contaminants may be found at the HPP Portal and PeptideAtlas.

### 4. FDR-related guidelines

- a. **Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels.** Describe which tools are used to estimate the false discovery rate (FDR) at the peptide-spectrum-match (PSM) level, at the distinct peptide sequence level, and at the protein level. Briefly describe the approach and what assumptions are made or implied, and any corrections for the fraction of the proteome covered. If you use novel or uncommon tools and criteria, compare your results with results with tools that are widely used in the community.
- b. **Report the PSM-, peptide-, and protein-level FDR values along with the total number of expected false positives at each level, using precision appropriate to the uncertainty in computed FDR.** Report the actual numbers of true positives and false positives at each level based on the thresholds used. Do not report the FDR with many significant digits since all current FDR calculation methods have substantial uncertainties.
- c. **Present large-scale results thresholded at equal to or lower than 1% protein-level global FDR, but not more than 10% local FDR.** The 1% is somewhat arbitrary but well accepted and remains set as the upper limit. As is frequently the case with datasets from modern instrumentation, a local FDR of 10% is reached before a 1% global FDR is reached, and then 10% local FDR threshold or less should be used instead. The common mistake of thresholding at a specific FDR and then proceeding as if all surviving results are correct, no matter how surprising, must be avoided.
- d. **If any large-scale datasets are individually thresholded and then combined, calculate the new, higher peptide- and protein-level FDRs for the combined result.** When datasets are combined, the true positives will mostly overlap, while the false positives will be scattered randomly across the proteome and thus overlap far less. This means that the FDR will be higher in the combined dataset.

Whereas the above guidelines apply generically to overall dataset analysis, the following guidelines apply specifically to the presentation of evidence of proteins that are not currently listed in the human reference proteome.

## 5. Guidelines for data-dependent acquisition (DDA) MS datasets

- a. **If using DDA mass spectrometry for such claims, present high mass-accuracy, high signal-to-noise ratio (SNR), and clearly annotated spectra to ensure PSMs are correct. Scrutinize spectra for missing and extra peaks.** Annotated spectra (i.e., spectra with the matched peaks clearly labeled) must be provided in the supplementary material for the manuscript. While low mass-accuracy and low SNR spectra can still be useful for many experiments, they are not acceptable for claims of new protein detections. Time-of-flight, FT-ICR, Astral, and Orbitrap-type instruments are considered in these guidelines as having high mass accuracy (when properly calibrated) in these guidelines. The spectra should be examined closely to determine if there are peaks missing that should be expected, if there are peaks present that are unexplained, and if a small alteration of the putative sequence would yield a much better match. This may indicate a false positive of a kind that is not modeled well by decoys.
- b. **Present high mass accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the spectra supporting the claims. Compute normalized backbone ion dot products between target spectra and synthetic spectra, as well as between target spectra and predicted spectra.** Synthetic peptides are powerful tools for determining the correct

identification of spectra. For each PSM corresponding to claim of a new PE1 protein, compare that PSM with a synthetic peptide (or recombinant protein product) spectrum of the same ion. All the major ions should have closely matching intensities in both spectra. If generating new reference spectra, it is encouraged to use the same high mass-accuracy instrument to verify matching intensity patterns and retention times. Closely matching spectra of the same peptidofrom ion (same modifications and charge) from SRMATlas, ProteomeTools, or similar resources is acceptable. Predicted spectra may be used in addition to or even in place of synthetic reference spectra. Compute normalized (0 to 1) dot product based on all backbone ions (typically b & y) (omitting b1 and y1) within acquired m/z range between the target spectra and reference spectra and report the results.

- c. **Provide Universal Spectrum Identifiers (USIs) for all natural and synthetic peptide spectra that support such claims, ideally as a supplementary data table.** The USI provides a mechanism to uniquely identify a spectrum being held up as evidence for an important claim. The USI will allow readers to access these important spectra in public data repositories in order to discuss correctness of the claims. See <http://psidev.info/USI> for more information.
6. **If using SRM verification for such claims, present target traces alongside synthetic heavy-labeled peptide traces, demonstrating co-elution and closely matching fragment mass intensity patterns.** All SRM runs performed must have spiked-in heavy labeled peptides corresponding to the putative identifications. The heavy-labeled peptides should be spiked in at an abundance similar to the target peptides so that minor impurities in the reference do not contribute to the target signal. Annotated chromatograms must be provided in the supplementary material of the manuscript. Solid peptide sequence evidence does not alter the uncertainties in matching that peptide uniquely to a protein (guideline 8). This guideline may also apply to PRM traces, although since PRM generates full MS/MS spectra, Guideline 5a-5c may be applied to PRM data instead. Guidelines 8 and 9 also apply for SRMs.
  7. **If using DIA MS, then, if the data are analyzed with XICs, apply the above SRM guidelines (6); if the data are analyzed by extracting deconvoluted spectra, apply the above DDA guidelines 5a-5c.** DIA-MS workflows such as SWATH-MS or the equivalent on other instrument types typically yield highly multiplexed spectra that make confident identification of peptides challenging. The guidelines that apply depend on the data analysis strategy. If the data are analyzed via extracted ion chromatograms (XICs) such as with DIA-NN, OpenSWATH, Spectronaut, PeakView, etc. then the SRM guideline 6 above applies. If the data are analyzed via extracted deconvoluted spectra such as with DIA-Umpire or DISCO, then the DDA Guideline 5a-5c above applies. In addition to the raw data, the extracted deconvoluted spectra must also be submitted to ProteomeXchange repository.
  8. **Even when very high confidence peptide identifications are demonstrated, consider alternate mappings of each peptide to canonical proteins other than the claimed one. Consider isobaric sequence/mass modification variants, all known SAAVs, and unreported SAAVs.** Even when a peptide identification is shown to be very highly confident, care should be taken when mapping it to a protein or novel coding element. Consider whether I=L, N[Deamidated]=D, Q[Deamidated]=E, GG=N, Q≈K, F≈M[Oxidation], or other isobaric or near

isobaric substitutions, or amino acid order inversions could change the mapping of the peptide from an extraordinary result to a mapping to a commonly-observed protein. Consider if a known single amino-acid variation (SAAV) in neXtProt could turn an extraordinary result into an ordinary result. Consider if a single amino-acid change, not yet annotated in a well-known source, could turn an extraordinary result into a questionable result. Check more than one reference proteome (e.g., RefSeq may have entries that UniProtKB and Ensembl do not, and vice versa). A tool to assist with this analysis is available at PeptideAtlas at <http://peptideatlas.org/map> (ProteoMapper).

9. **Support Tier 1 (1A digest, 1B HLA) claims via two or more distinct uniquely-mapping, non-nested peptide sequences of length  $\geq 9$  amino acids with the above evidence in the same paper. When 2 peptides overlap, the total extent must be  $\geq 18$  amino acids. 100% coverage, except for an initiating methionine, may be considered Tier 1 evidence. Tier 2A (digest) or Tier 2B (HLA) evidence may constitute just a single uniquely-mapping peptide of length  $\geq 9$  amino acids. Tier 3 evidence (not meeting Tier 2A or 2B evidence) may be discussed but is generally not considered sufficient for a claim of detection.** Single-peptide detections have too high a chance of being some type of pernicious false positive to be sufficient for claiming a new protein detection at the highest confidence. Likewise, short peptides of length 8 or smaller have relatively few peaks and have an increased chance of mapping to immunoglobulins or other variable sequences not readily apparent in the reference proteome. Nested peptides (where one sequence is fully subsumed within another) do provide additional confidence that the peptide identification is correct, but provide no additional evidence that the peptide-to-protein mapping is unique. However, microproteins can be very short, and optimal evidence can be challenging to achieve. Therefore, a tiered system of detection claims is provided. Tier 1 claims require at least 2 uniquely mapping peptides of 9 residues or longer that together cover an extent of at least 18 amino acids. For very short microproteins, 100% coverage (with the possible exclusion of the initiating methionine) can be considered Tier 1 evidence. When Tier 1 evidence is achieved via digest data (e.g. trypsin or other proteases), it is considered Tier 1A. When Tier 1 evidence is achieved in whole or in part via immunopeptidome enrichment data, it is considered Tier 1B. Lesser evidence that includes at least 1 uniquely mapping peptide of 9 residues or longer is considered Tier 2A (digest) or Tier 2B (immunopeptidome). Evidence that does not meet the above criteria may be termed Tier 3, but is generally not considered sufficient for a claim of detection. Nonetheless, such “candidate detections” may be discussed to enable capture of this information by other researchers for follow up by further experiments.