Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*J Proteome Res*. Author manuscript; available in PMC 2021 July 06.

Published in final edited form as:

*J Proteome Res*. 2021 April 02; 20(4): 1986–1996. doi:10.1021/acs.jproteome.0c00799.

# Enhancing Open Modification Searches via a Combined Approach Facilitated by Ursgal

**Stefan Schulze**[#1,*], **Aime Bienfait Igiraneza**[#1], **Manuel Kösters**[2], **Johannes Leufken**[2], **Sebastian A. Leidel**[2], **Benjamin A. Garcia**[3], **Christian Fufezan**[4], **Mechthild Pohlschröder**[1,*]

[1]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA [2]Department of Chemistry and Biochemistry, University of Bern, 3012 Bern, Switzerland [3]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA [4]Institute of Pharmacy and Molecular Biotechnology, Heidelberg University, 69120 Heidelberg, Germany

[#] These authors contributed equally to this work.

## Abstract

The identification of peptide sequences and their post-translational modifications (PTMs) is a crucial step in the analysis of bottom-up proteomics data. The recent development of open modification search (OMS) engines allows virtually all PTMs to be searched for. This not only increases the number of spectra that can be matched to peptides but also greatly advances the understanding of biological roles of PTMs through the identification, and thereby facilitated quantification, of peptidoforms (peptide sequences and their potential PTMs). While the benefits of combining results from multiple protein database search engines has been established previously, similar approaches for OMS results are missing so far. Here, we compare and combine results from three different OMS engines, demonstrating an increase in peptide spectrum matches of 8–18%. The unification of search results furthermore allows for the combined downstream processing of search results, including the mapping to potential PTMs. Finally, we test for the ability of OMS engines to identify glycosylated peptides. The implementation of these engines in the Python framework Ursgal facilitates the straightforward application of OMS with unified

[*]**Corresponding Authors**: Mechthild Pohlschröder – Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA, pohlschr@sas.upenn.edu, Stefan Schulze – Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA, sschulze@sas.upenn.edu.

Author Contributions

S.S. and A.B.I. contributed equally to this work. S.S. and A.B.I. performed the analysis of the datasets and their results. S.S., A.B.I., M.K., J.L., and C.F. developed Ursgal and implemented new algorithms. Results were interpreted and the manuscript was written through contributions of all authors. M.P., S.A.L. and B.A.G provided funding and access to MS data and computational resources. S.S. and M.P. developed the idea for the project and supervised it. All authors have given approval to the final version of the manuscript.

Supporting Information.

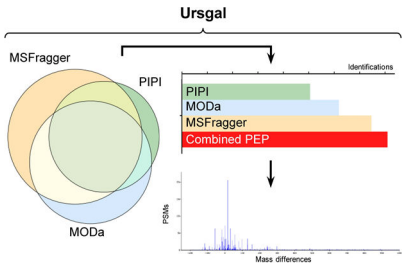
Figure S1: Effects of sanitizing results from OMS engines; Figure S2: Overlap between different OMS engines for results from prokaryotic proteome datasets; Figure S3: Increase of commonly identified spectra through the combination of results from different OMS engines; Figure S4: PSM differences between OMS engines for results from prokaryotic proteome datasets; Figure S5: Mass difference profiles for *H. volcanii*; Figure S6: Mass difference profiles for *E. coli*; Figure S7: Comparisons between OMS and CS results; Figure S8: Identification of O-glycopeptides through the OMS approach.

The authors declare no conflict of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

parameters and results files, thereby enabling yet unmatched high-throughput, large-scale data analysis.

## Graphical Abstract




### Keywords

Proteomics; Open modification search; Python; Bioinformatics; Glycosylation; Post-translational modifications

## INTRODUCTION

After transcription of a gene and translation of mRNA, the synthesized protein can undergo a variety of modifications. These post-translational modifications (PTMs) result in a multitude of variations, also called proteoforms, produced from a single gene. These proteoforms can differ in their structure, localization and enzymatic activity, thus increasing complexity from genome to proteome level[1–5]. PTMs often allow for subpopulations of the proteome to be changed within seconds by simple attachment or detachment of specific PTMs. As a result, intricate networks of PTMs are involved in the regulation of virtually all biological processes, in all three domains of life[6]. For instance phosphorylation, one of the best-studied PTMs, is the driving force of many signaling cascades[7–9]. In contrast to this simple phosphate modification, protein glycosylation is one of the most complex PTMs[10]. It has been shown to be involved in a multitude of cellular processes[11,12], including the formation of biofilms, microbial communities crucial for virulence and antimicrobial resistance of various pathogens[13,14]. Many more PTMs are known to be correlated with a variety of human diseases[15], including a complex system of histone modifications[16]. Thus, gaining a better understanding of the biological roles and regulation of PTMs can not only provide deeper insights into cell biology but is also of great biomedical relevance. However, the complexity of PTMs, their often low abundance as well as their non-template-driven biosynthesis makes their analysis challenging.

Mass spectrometry is the method of choice for studying whole proteomes and related PTMs. Commonly, there are two approaches, namely top-down and bottom-up proteomics, for the analysis of whole proteins and peptides resulting from proteolytic digestion, respectively. The identification of intact proteoforms in top-down proteomics provides unique opportunities, such as insights into relationships between PTMs on the same protein[17]. However, technical challenges, both in the mass spectrometric measurement of intact proteoforms and the downstream analysis of resulting spectra, have limited the large-scale

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

application of top-down proteomics so far. Bottom-up proteomics, on the other hand, has been widely applied due to its relatively simple setup, resulting in the broad availability of bioinformatic tools. In order to identify peptides from bottom-up proteomics data, most commonly a protein database search is performed, comparing precursor mass (MS1 level) and peptide fragmentation spectra (MS2 level) to theoretical spectra derived from *in silico* digests of a protein database. For the identification of PTMs, different search engines allow for either a closed or an open modification search to be performed.

In a closed search (CS), a limited number of modifications is defined and during the search the masses of the modifications are added to each potential modified residue. This addition of variable modifications in a protein database search exponentially increases the search space, since each peptide with modifiable residues is searched in all possible combinations, i.e. including all peptidoforms. Therefore, the search for all potential modifications has long been virtually impossible, due to the lack of optimized computational resources[18]. In contrast, an open modification search (OMS) allows to search for all modifications within a user-defined mass range, e.g. ±500 Da. In its simplest form, the precursor mass tolerance in protein database searches can be increased to the desired mass range and the mass of peptide modifications is determined as the mass difference between precursor and peptide mass[19]. However, this approach does not account for fragment ions that are shifted as a result of a modification.

Recently, a variety of dedicated OMS engines such as MODa[20], PIPI[21], MSFragger[22] and TagGraph[23] have been developed, increasing the speed and accuracy of this approach, and taking into account shifted fragment ions through different search strategies. MODa, PIPI and TagGraph employ a strategy in which sequence tags (short substrings of a peptide sequence) are matched to measured spectra (TagGraph uses de novo search results for this); multiple matching tags are then aligned and the delta masses between the tags and/or between the sum of the tags and the precursor mass is reported as mass(es) of the modification(s). In contrast, MSFragger generates a fragment ion index that is used for the matching of peaks in measured spectra. Through a recent update of the algorithm, shifted ion indexes can be generated as well, allowing for the assignment of modified fragment ions as well as the localization of modifications[24]. These varied approaches have greatly advanced OMSs. Nevertheless, as for CSs, the choice of the most suitable OMS engine is left to the user and since each OMS engine comes with its own advantages and disadvantages, the decision becomes difficult.

Additionally, taking advantage of the diversity in search algorithms, the combination of results from multiple search engines has been shown to increase the number of identifications for CS approaches[25,26]. At the same time, combining results from different search engines can be used to re-evaluate false discovery rates (FDRs) for peptide spectrum matches (PSMs) and thereby increase their reliability[25–28]. Different methods have been developed for this task, including machine learning approaches[29]. In general, this builds on the intuitive assumption that PSMs independently identified by multiple tools tend to be more reliable than PSMs only identified by a single engine. However, the approach to combine results from different search engines has not been applied to OMSs yet.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In this work, we provide a unified scriptable access to various OMS engines. We compare the results from different OMS engines in regard to their level of disagreement on PSMs for the same spectra. Furthermore, we show that combining results from different engines increases the number of peptide identifications and can aid in the identification of PTMs. Finally, comparisons with traditional CS approaches reveal large overlaps, indicating the potential to use OMS as a standard search approach. However, taking the example of glycosylation, we also reveal limitations of the OMS approach.

## MATERIAL AND METHODS

### Implementation of OMS engines in Ursgal

The Python framework Ursgal has previously been described in detail[26] and an extensive, continuously updated documentation is available (https://ursgal.readthedocs.io/en/latest). The OMS engines MODa (v. 1.61)[20], PIPI (v. 1.4.6)[21], MSFragger (v. 2.3)[22] and TagGraph[23] have been implemented as protein database search engines. The tools PTM-Shepherd[30] and PTMiner[31] have been included for the downstream processing of mass differences reported by OMS engines. All relevant parameters to execute these engines have been unified and are available within Ursgal's uparams. An overview of all parameters, engine specific parameters, parameter translations between engines, and more can be found in the documentation as well as through an interactive, searchable Dash app user interface.

Results from each tool are converted into a unified comma separated values file (CSV) format, in which each row contains a PSM, corresponding to a spectrum, and all properties of the PSM are listed in distinct columns (e.g. "Spectrum Title", including the file name and spectrum ID; "Sequence", referring to the peptide sequence; "Modifications", PTMs given as PSI-MS terms together with their position and separated by semicolons; "Protein ID"; engine scores, etc). For the results of OMS engines, this format has been extended by a "Mass Difference" column, comprising the mass difference(s) reported by the engine in Da. PTMs assigned by downstream processing tools are included in the column "Mass Difference Annotations". If a PTM could not be assigned, the mass difference is updated to the corresponding binned mass difference peak.

Further details, including example scripts for the use of OMS and downstream processing engines can be found in the Ursgal GitHub repository (https://github.com/ursgal/ursgal). The current version of Ursgal has been uploaded to Zenodo with the following permanent digital object identifier https://doi.org/10.5281/zenodo.4299358.

### Datasets and protein databases

Four datasets were used: two *Homo sapiens* datasets, one *Haloferax volcanii* dataset and one *Escherichia coli* dataset. The first *H. sapiens* dataset (PXD004452), an in-depth proteomics dataset, was published by Bekker-Jensen et al[32]. The second *H. sapiens* dataset (PXD013715), which is a glycoproteomic dataset, was published by Brown et al.[33] The *E. coli* dataset (PXD000498) was published by Schmidt et al[34]. Details on the sample preparation and mass spectrometric measurements for all these datasets can be found in the original publications. Finally, the *H. volcanii* dataset represents a glycoproteomic analysis,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with details about the sample preparation and analysis of results given in the PRIDE description (PXD021874). Raw files were converted into mzML using msConvert within the ProteoWizard Toolkit[35] (v. 3.0.19046).

Protein databases for *H. sapiens* (UP000005640) and *E. coli* (UP000000625) were downloaded from UniProt on November 18, 2019 and June 13, 2020, respectively. The *H. volcanii* (https://doi.org/10.5281/zenodo.3565631) reference proteome was acquired from the Archaeal Proteome Project[36]. The target decoy databases were generated the same way for all three organisms: the reference databases were first supplemented with the cRAP database (https://www.thegpm.org/crap/) of common contaminants before generating decoys by peptide shuffling using the generate_target_decoy_1_0_0 node within Ursgal. Different decoys were generated depending on the enzyme used for the respective datasets (Trypsin or GluC).

## Open modification search

The OMS pipeline was set up within Ursgal (v. 0.6.7)[26] and included the conversion of mzML files to mgf format using pymzML (v. 2.4.6)[37] before the main search was carried out by three search engines, namely MODa (v. 1.61)[20], PIPI (v. 1.4.6)[21] and MSFragger (v. 2.3)[22]. If not defined otherwise, Ursgal default values within the profile QExactive+ were used. Carbamidomethylation of C was set as the fixed modification and the enzyme for *in silico* digestion was set as either trypsin or gluc depending on the input file. Precursor and fragment mass tolerances as well as modification sizes used for each dataset are summarized in Table 1. The parameter moda_high_res was set to False. Mapping of peptide sequences to proteins was done using the upeptide_mapper_1_0_0 node.

## Closed search

The CS pipeline was set up within Ursgal[26] (v. 0.6.7) similarly to the OMS pipeline. The following search engines were used: X! Tandem[38] (v. Vengeance), MSFragger[22] (v. 2.3), and MS-GF+[39] (v. 2019.07.03). In addition to setting the fixed modification as Carbamidomethylation of C, the following variable modifications were used: methionine oxidation and N-terminal acetylation. The remaining parameters and processing steps were similar to those of the OMS except that the modifications were only allowed within the range of the precursor mass tolerance.

## Statistical post-processing

Results from fractions of the same sample were merged before post-processing with Percolator[40] (v. 3.4) to determine PEPs. For individual search engine results, validated results were filtered by a PEP ≤ 1%. For the combined PEP approach[26], unfiltered but validated results of all employed engines were subjected to the combine_pep_1_0_0 node and results were subsequently filtered by a combined PEP ≤ 1%. In order to remove conflicting PSMs for the same spectra, sanitizing was performed using the corresponding Ursgal node. PSMs were ranked based on their combined PEP and only the best scoring PSM per spectrum was accepted. Given that some engines report multiple PSMs per spectrum, individual search engine results were also sanitized as described above, using the PEP instead of combined PEP. To merge rows corresponding to the same PSM, minimum

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

values of combined PEP and Bayes PEP were maintained in the case of the combined PEP approach, while minimum values for PEP were considered for the individual PEP approach.

### Analyses of mass differences

Mass differences identified during OMS were analyzed using PTM-Shepherd[30] (v. 0.3.5) which was implemented in Ursgal. Results from either single OMS engine analyses, filtered by 1% PEP and sanitized, or combined results from multiple OMS engines, filtered by 1% combined PEP and sanitized, were used as input files for PTM-Shepherd. Parameters for PTM-Shepherd runs were based on parameters used in the OMSs for each dataset. In addition to that, the bin size for mass differences was set to 0.2 mDa, and the minimum relative peak intensity was set to 0.01. It should be noted that PTM-Shepherd allows only one mass difference per PSM, which is why multiple mass differences reported (e.g. by MODa) were summed up into one combined mass difference. Furthermore, MODa reports mass differences as whole integers instead of floats as the other employed OMS engines. For the analysis of mass differences from only MODa results, these values were used, whereas for the analysis of combined OMS results, mass differences were recalculated as the difference between the precursor mass and the mass of the peptide (including fixed modifications).

### Comparisons between OMS engine results

Results of different OMS engines were compared using the Ursgal venndiagram_1_1_0 node. Comparisons between open search engines' identifications were performed at the peptide level and the spectrum level. To further check for disagreements between engines, spectra identified by different engines were analyzed for corresponding peptide assignments of each engine and disagreements were further investigated by comparing amino acids as well as peptide lengths.

### Comparisons between OMS and CS

Combined results from all OMS engines were compared to combined results from all CS engines in regard to peptide and spectra identifications. Equivalent to the comparison of results from different engines in OMS, peptides mapped to the same spectra by both searches were compared. Disagreements were further investigated by comparing amino acids as well as peptides' lengths. In addition, glycopeptides and corresponding spectra identified through CS (including through the use of glycopeptide-focused search engines) were compared to open search results that were processed as described below.

### Glycoproteomic analyses

For the glycoproteomic analysis of human datasets, pGlyco[41] (v. 2.0) was employed as implemented in Ursgal, which can be considered a CS using a glycan database in addition to a protein database. Parameters were based on those used in the CS approach (see above), and default values were used if not indicated otherwise. Glycopeptides matched by pGlyco were statistically post-processed using the pGlyco-internal FDR estimation algorithm. Results were filtered by a q-value ≤ 1%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The relatively short, linear *N*-glycans of *H. volcanii* allowed for them to be included as potential modifications in CSs using the same engines and parameters as described above. The included potential modifications were the only differences and contained oligosaccharides for each step of the AglB- and Agl15-dependent *N*-glycosylation pathways[42] as well as two hexoses as *O*-glycans. Since some engines do not allow multiple modifications for the same amino acid, each oligosaccharide was searched for in a separate run. Results were then combined after individual statistical post-processing with Percolator for each modification. Glycopeptide PSMs were selected as PSMs containing a modification with at least one hexose.

For the OMS approach, mass differences were post-processed with PTM-Shepherd (see above). Afterwards, mass differences were mapped to glycan masses from a human glycan database, downloaded from glySpace using GlycReSoft[43], for the human dataset and to glycan masses used as modifications in the CS for the *H. volcanii* dataset. For matches between reported mass differences and glycan masses a mass tolerance corresponding to the used precursor mass tolerance was allowed.

## RESULTS

OMSs have become increasingly popular for the analysis of bottom-up proteomics data. Through the integration of various OMS engines in Ursgal[26], we provide a scriptable interface to these tools within the widely used Python environment (https://github.com/ursgal/ursgal). The unification of search parameters and output formats ensures that complex workflows can be easily generated and facilitates the straightforward comparison and downstream processing of results.

### The combination of different OMS engines increases the number of identified peptides.

The combination of results from multiple search engines has previously been shown to be beneficial for the number of identifications as well as their reliability[25–29]. However, this approach has not been applied to OMSs yet. Therefore, we have analyzed an in-depth human proteomics dataset (PXD004452)[44], using three OMS engines (MODa, PIPI and MSFragger) to evaluate the applicability of this approach (Fig. 1).

The comparison of the results for each search engine, filtered by 1% posterior error probability (PEP), showed a higher number of peptide sequence identifications for MSFragger (149,426) than for MODa (123,765) and PIPI (101,063). While the majority of peptide sequences was identified by multiple engines (62%), the substantial number of unique identifications by each engine indicated that a combination of results could be beneficial (Fig. 1a). However, the estimation of PEPs for each individual engine's results prevents a simple merge of the identifications, since this would lead to an accumulation of false positive identifications. Therefore, in order to combine results from different search engines, while controlling the PEP at the same time, a combined PEP approach was employed as described previously[26]. As expected, the overlap of identified peptide sequences between the engines increased substantially (Fig. 1b). Furthermore, the number of identified peptides increased for each individual engine, most strikingly for MODa and PIPI. This indicates that the combination of multiple search engine results, and the concomitant

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

re-ranking of PSMs, helps to recover PSMs with PEPs > 1% for their individual search engine results. However, in order to compare the total number of identifications from the combined PEP approach with identifications from single search engines, results were sanitized to remove conflicting PSMs from different engines. While this removed around 15,000 peptide sequences (Fig. S1), the remaining total of 162,019 identified peptide sequences represents an increase of 8.4% in comparison to the best single search engine, i.e MSFragger (Fig. 1c). For comprehensive datasets from *Haloferax volcanii* (PXD021874) and *Escherichia coli* (PXD000498)[34] (Fig. S2), both of which have less complex proteomes than *Homo sapiens*, the same trend in regard to the total number of peptide identifications was observed, with an increase of 15.0% and 17.9%, respectively (Fig. 1c).

It should be noted that the number of spectra with corresponding PSMs slightly decreased by 3.5% in the human dataset employing the combined PEP approach in comparison to the best performing search engine (Fig. S3). Lower identification rates by single engines, e.g. PIPI identified 50% less PSMs than MSFragger, might have contributed to this effect, since the overlap between the OMS engines was limited by the engine with the lowest identification rate. However, for the *H. volcanii* and *E. coli* datasets, the total number of identified spectra increased by 6.9% and 3.1%, respectively. For all datasets, the number of overlapping spectra between all engines increased substantially, indicating a higher degree of agreement between the engines (see below).

## Differences in PSM assignments between search engines are largely attributed to differences in peptide length

While the comparison of results from different search engines is often focused on the number of identified peptides or PSMs[24], the level of (dis-)agreement on PSMs between various tools is rarely analyzed. However, this is of special interest for the combination of multiple OMS results, since PSM agreements are considered more reliable than PSMs that show disagreements between search engines. Therefore, we have determined how often engines map the same spectra to the same peptide sequence, i.e. how often they agree on PSMs. When considering 1% PEP-filtered PSMs from individual search engine analyses of the human dataset, 84% of all spectra that resulted in peptide identifications for all three search engines, showed agreements in the corresponding PSMs (Fig. 2a). This is similar to the level of agreement for the combined PEP results (82%), however, the number of overlapping spectra, and therefore the total number of PSM agreements, increased substantially by roughly 60% (Fig. 2b). The number of spectra that showed an overlap between two search engines was comparable in both approaches, but the percentage of PSM agreements increased from 75% to 84% for the individual and combined PEP approach, respectively (Fig. 2c,d).

Importantly, the vast majority of differences in PSM assignments between search engines comprised length variations of the matched peptide (with the shorter peptide being a full substring of the longer peptide) (Fig. 2). This indicates that part of the mass differences (of shorter peptides) were attributed to additional amino acids. While, as a result, PSMs for these spectra differed, they did not represent clear disagreements between the engines, especially since the OMS engines employed here reported mass differences rather than

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

defined PTMs. These mass differences could subsequently be mapped to additional (or fewer) amino acids and/or PTMs. Similarly, differences in ≤ 3 amino acids could be explained by mass shifts due to modifications like amidation/deamidation that lead the conversion of amino acids. Besides these differences in PSM assignments, only 1% to 3% of PSMs were associated with completely different peptide identifications between the search engines, representing clear disagreements between the engines. Similar trends were observed for the *H. volcanii* and *E. coli* datasets, however, with an overall lower level of PSM differences (Fig. S4). While the observed ratio of PSM disagreements was close to the 1% PEP threshold, it should be noted that this does not allow for conclusions about the overall PEP, since only a subset of PSMs was taken into account (spectra with PSM assignments by multiple engines). Furthermore, PSM disagreements could also indicate chimeric spectra instead of false assignments by either engine.

For further downstream processing of OMS results, differing PSMs for single spectra were reconciled using the Ursgal sanitize node. PSMs of one spectrum were ranked according to their score (combined PEP) and only the best scoring PSM was accepted. The combined PEP approach is beneficial to this process, since the number of engines that assign the same PSM affects the final score that is used for the ranking of differing PSMs.

## Unified OMS results facilitate the combined post-processing of mass differences

Mass differences can be mapped to known PTMs in order to gain more information about the corresponding type of the modification. This post-processing includes the fitting of mass difference profiles and their matching to databases of known PTMs[30,31,45,46]. Recently, several tools have been developed for this purpose, including PTM-Shepherd[30] and PTMiner[31], both of which have been implemented in Ursgal, allowing for the unified post-processing of combined OMS results. Using PTM-Shepherd as an example, we observed that mass difference profiles varied substantially between the different OMS engines (Fig. 3). PIPI identified a comparatively small range and overall lower number of mass differences. MODa reports mass differences only as integers, which limits its use for PTM mapping. For the combined OMS results, mass differences of MODa results were therefore recalculated as the difference between the precursor mass and the peptide mass (including fixed modifications). The mass difference profile of the combined OMS results is most similar to the one of MSFragger. Nevertheless, some mass differences, e.g. 42.01 Da (acetylation) and 79.97 (phosphorylation), were more prominently observed in results from MODa and/or PIPI and are therefore more abundant in the combined OMS results as well. For the datasets of *H. volcanii* and *E. coli*, the mass difference profile is less complex (Fig. S5 and S6), which is in line with the lower complexity of the respective proteomes. Nevertheless, for PIPI a lack of mass differences in the higher mass range is confirmed in the *H. volcanii* dataset.

## OMS and CS lead to complementing results.

OMSs are often performed in conjunction with CSs, e.g. as cascaded searches in which unidentified spectra in a CS are subsequently analyzed with an OMS[47]. However, increasing evidence suggests that OMSs can reliably identify unmodified peptides as well, questioning the need for cascaded search approaches or even CSs in general[22,24]. Therefore, we

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

compared the results from the combined OMS approach to results from a combined CS approach. In both cases, three search engines were used and results were filtered by 1% combined PEP. For the human proteome dataset, the OMS approach identified almost 30% more peptide sequences in 72% more spectra (Fig. 4). A large overlap between both approaches on the peptide and spectrum level, as well as a high level of agreement on PSMs for the same spectra, indicated that unmodified peptides are indeed readily identified by the OMS approach. Even when taking into account only modifications that are commonly included in CSs (oxidation of M, acetylation of protein N-termini), the OMS approach identified more than twice as many unique peptide sequences as the CS approach. Nevertheless, unique identifications from the CS approach increased the total number of identified peptides by 12%. Similar trends were observed for the *H. volcanii* and *E. coli* datasets (Fig. S7).

Interestingly, a large part of the PSM differences between the OMS and CS approach was due to the identification of longer peptide sequences in the CS (Fig. 4c). Surprisingly, the vast majority of shorter peptides assigned by OMS engines (MSFragger and MODa) harbored mass differences that were attributed to an addition of K or R by PTM-Shepherd. This means that instead of matching a sequence with a missed cleavage site (within the limit of maximum missed cleavage sites), MSFragger and MODa frequently added the additional amino acid as a mass difference. This effect is reduced in the combined PEP results (Fig. 3, insets), highlighting another advantage of this approach since a matching behavior like this is in most cases not desired because it would complicate the interpretation of additional modifications on peptides with (potential) missed cleavage sites.

## Complex glycopeptide identifications remain challenging for OMS engines

Recent studies suggested that OMS engines can be used for the identification of glycosylated peptides and/or the generation of glycan databases for subsequent glycopeptide-specific searches[48–51]. We therefore compared the combined OMS results to glycopeptide identifications from the established, specialized search engine pGlyco (v. 2.0)[41], which uses a glycan database and could therefore be seen as a glycopeptide-centric CS approach. Since the regular human proteome dataset yielded only a few glycopeptide identifications (Fig. 5a,d), we analyzed a human dataset that focused on *N*-glycoproteomics[33] (Fig. 5b,e). While OMS engines could identify a large number of *N*-glycopeptides, only 50% of the spectra identified by the glycopeptide-centric CS were included in the OMS results. In addition, for only a subset (710 out of 1450) of these spectra, the mass shifts identified in the OMS approach could be matched to *N*-glycans. While this subset of spectra showed a high level of PSM agreement (>98%) between the open modification and CS results, a closer look into the remaining spectra revealed differences in the identified peptide sequence for 49% of the spectra (Fig. 5e). For the majority of these differences, the OMS engines matched longer peptide sequences to the spectra than the CS engine. This means that mass differences were attributed to additional amino acids by the OMS but to *N*-glycans by the glycopeptide-centric CS. Interestingly, a much lower level of PSM differences was observed for the *H. volcanii* dataset (Fig. 5c,f). This could be explained by the fact that *H. volcanii* synthesizes linear *N*-glycans, which comprise four to five monosaccharides[42] and are therefore smaller and less complex than

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

human *N*-glycans. While, due to variations in the fragmentation behavior of different types of glycosylation, we differentiated between *N*- and *O*-glycopeptides in our analysis (Fig. 5 and Fig. S8, respectively), the same tendencies were observed regardless of the glycan type.

## DISCUSSION

Over the last years, OMS engines have matured into invaluable tools for comprehensive bottom-up proteomics, facilitating the analysis of a broad variety of PTMs. This work now provides a unified, scriptable access to multiple OMS engines through their integration into the Python framework Ursgal. Comparisons between MSFragger, MODa and PIPI showed large overlaps and only little PSM disagreements between the engines. These results provided the basis for the application of a combined PEP approach, which demonstrated its usefulness by increasing the number of peptide identifications by 8–18%. Furthermore, the unified and combined results of all OMS results can be post-processed using PTM-Shepherd or PTMiner, also included in Ursgal, facilitating the mapping of mass differences to known PTMs. The use of additional OMS engines could lead to further increases in identification rates, especially if tools with complementing search algorithms are used. For example, TagGraph performs an OMS using *de novo* search results[23], is implemented in Ursgal, and could benefit from the availability of multiple *de novo* search engines within Ursgal. However, the unique scoring and statistical post-processing algorithm employed by TagGraph complicates its integration in a combined PEP approach. Nevertheless, the modular structure of Ursgal allows for the straightforward implementation of additional tools, e.g. search engines like MetaMorpheus[52] or post-processing tools like Crystal-C[53].

Comparisons with CS results showed a large overlap and small number of PSM disagreements between open modification and CS results. While these results indicate a potential for the replacement of CS approaches with OMSs, especially since the latter led to a higher number of identifications, employing both approaches still provides the most comprehensive results. The unified, scriptable access to both, open modification and CS engines within Ursgal allows for the straightforward generation of workflows taking advantage of both approaches.

Furthermore, we showed limitations in the identification of complex modifications like glycosylation. OMSs showed a lower glycopeptide identification rate in comparison to glycopeptide-centric CSs and a high degree of PSM differences was observed between the two approaches. This demonstrates the usefulness of specialized search engines, which, as in the case of pGlyco, include the search for glycopeptide-specific fragment ions like oxonium- and Y-ions. This is in line with the recent development of glycopeptide-centric versions of OMS engines, e.g. MSFragger-Glyco[54] and MetaMorpheus O-Pair Search[50], the former of which allows searching for oxonium- and Y-ions. However, these tools and search modes take advantage of glycan databases, which are not comprehensively available for all types of glycosylation or all organisms. Therefore, it had been suggested, and successfully applied, that explorative OMSs could be used to identify present glycan masses and thereby generate glycan databases for the respective dataset/organism[48,51]. The differences between OMS and glycopeptide-specific CSs that we observed in the example datasets here are less likely to affect this approach. Even though OMS engines tended to attribute part of the glycan mass

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to additional amino acids, alternative mass differences did not match to glycan masses and would therefore not be added to a glycan database. Furthermore, glycan compositions potentially falsely added to a glycan database would undergo additional scrutiny in subsequent glycopeptide-centric searches and therefore do not necessarily lead to false glycopeptide identifications. Nevertheless, a more sophisticated post-processing for potential glycopeptide identifications from OMSs seems to be required to take full advantage of OMS engines for the generation of glycan databases as well as for the direct and reliable identification of glycopeptides.

In conclusion, we demonstrated the feasibility and advantages of combining results from multiple OMS engines as well as some limitations of this approach. The scriptable interface, simple extensibility and open access of Ursgal facilitates the broad application of this approach as well as the integration of future improvements.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Funding Sources

S.S. was funded by the German Research Foundation (DFG Postdoctoral Fellowship, 398625447). A.B.I was supported by the Spring 2018 Pincus-Magaziner Family Undergraduate Research and Travel Fund from the College Alumni Society and the Seltzer Family Digital Media Award at the University of Pennsylvania. M.P., A.B.I. and S.S. were supported by the National Science Foundation Grant 1817518. MK, JL, SAL were funded by the NCCR RNA & Disease (Swiss National Science Foundation).

## REFERENCES

(1). Aebersold R; Agar JN; Amster IJ; Baker MS; Bertozzi CR; Boja ES; Costello CE; Cravatt BF; Fenselau C; Garcia BA; Ge Y; Gunawardena J; Hendrickson RC; Hergenrother PJ; Huber CG; Ivanov AR; Jensen ON; Jewett MC; Kelleher NL; Kiessling LL; Krogan NJ; Larsen MR; Loo JA; Loo RRO; Lundberg E; MacCoss MJ; Mallick P; Mootha VK; Mrksich M; Muir TW; Patrie SM; Pesavento JJ; Pitteri SJ; Rodriguez H; Saghatelian A; Sandoval W; Schlüter H; Sechi S; Slavoff SA; Smith LM; Snyder MP; Thomas PM; Uhlén M; Van Eyk JE; Vidal M; Walt DR; White FM; Williams ER; Wohlschlager T; Wysocki VH; Yates NA; Young NL; Zhang B How Many Human Proteoforms Are There? Nat. Chem. Biol 2018, 14 (3), 206–214. 10.1038/nchembio.2576. [PubMed: 29443976]

(2). Spoel SH Orchestrating the Proteome with Post-Translational Modifications. J. Exp. Bot 2018, 69 (19), 4499–4503. 10.1093/jxb/ery295. [PubMed: 30169870]

(3). Bludau I; Aebersold R Proteomic and Interactomic Insights into the Molecular Basis of Cell Functional Diversity. Nat. Rev. Mol. Cell Biol 2020, 21 (6), 327–340. 10.1038/s41580-020-0231-2. [PubMed: 32235894]

(4). Smith LM; Kelleher NL Proteoforms as the next Proteomics Currency. Science 2018, 359 (6380), 1106–1107. 10.1126/science.aat1884. [PubMed: 29590032]

(5). Smith LM; Kelleher NL Proteoform: A Single Term Describing Protein Complexity. Nat. Methods 2013, 10 (3), 186–187. 10.1038/nmeth.2369. [PubMed: 23443629]

(6). Barber KW; Rinehart J The ABCs of PTMs. Nat. Chem. Biol 2018, 14 (3), 188–192. 10.1038/nchembio.2572. [PubMed: 29443972]

(7). Denhardt DT Signal-Transducing Protein Phosphorylation Cascades Mediated by Ras/Rho Proteins in the Mammalian Cell: The Potential for Multiplex Signalling. Biochem. J 1996, 318 (Pt 3), 729–747. [PubMed: 8836113]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(8). Ardito F; Giuliani M; Perrone D; Troiano G; Muzio LL The Crucial Role of Protein Phosphorylation in Cell Signaling and Its Use as Targeted Therapy (Review). Int. J. Mol. Med 2017, 40 (2), 271–280. 10.3892/ijmm.2017.3036. [PubMed: 28656226]

(9). Day EK; Sosale NG; Lazzara MJ Cell Signaling Regulation by Protein Phosphorylation: A Multivariate, Heterogeneous, and Context-Dependent Process. Curr. Opin. Biotechnol 2016, 40, 185–192. 10.1016/j.copbio.2016.06.005. [PubMed: 27393828]

(10). Hart GW Glycosylation. Curr. Opin. Cell Biol 1992, 4 (6), 1017–1023. 10.1016/0955-0674(92)90134-X. [PubMed: 1485955]

(11). Varki A Biological Roles of Glycans. Glycobiology 2017, 27 (1), 3–49. 10.1093/glycob/cww086. [PubMed: 27558841]

(12). Reily C; Stewart TJ; Renfrow MB; Novak J Glycosylation in Health and Disease. Nat. Rev. Nephrol 2019, 15 (6), 346–366. 10.1038/s41581-019-0129-4. [PubMed: 30858582]

(13). Schäffer C; Messner P Emerging Facets of Prokaryotic Glycosylation. FEMS Microbiol. Rev 2017, 41 (1), 49–91. 10.1093/femsre/fuw036. [PubMed: 27566466]

(14). Esquivel RN; Schulze S; Xu R; Hippler M; Pohlschroder M Identification of Haloferax Volcanii Pilin N-Glycans with Diverse Roles in Pilus Biosynthesis, Adhesion, and Microcolony Formation. J. Biol. Chem 2016, 291 (20), 10602–10614. 10.1074/jbc.M115.693556. [PubMed: 26966177]

(15). Xu H; Wang Y; Lin S; Deng W; Peng D; Cui Q; Xue Y PTMD: A Database of Human Disease-Associated Post-Translational Modifications. Genomics Proteomics Bioinformatics 2018, 16 (4), 244–251. 10.1016/j.gpb.2018.06.004. [PubMed: 30244175]

(16). Britton L-MP; Gonzales-Cope M; Zee BM; Garcia BA Breaking the Histone Code with Quantitative Mass Spectrometry. Expert Rev. Proteomics 2011, 8 (5), 631–643. 10.1586/epr.11.47. [PubMed: 21999833]

(17). Catherman AD; Skinner OS; Kelleher NL Top Down Proteomics: Facts and Perspectives. Biochem. Biophys. Res. Commun 2014, 445 (4), 683–693. 10.1016/j.bbrc.2014.02.041. [PubMed: 24556311]

(18). Ahrné E; Müller M; Lisacek F Unrestricted Identification of Modified Proteins Using MS/MS. PROTEOMICS 2010, 10 (4), 671–686. 10.1002/pmic.200900502. [PubMed: 20029840]

(19). Chick JM; Kolippakkam D; Nusinow DP; Zhai B; Rad R; Huttlin EL; Gygi SP An Ultra-Tolerant Database Search Reveals That a Myriad of Modified Peptides Contributes to Unassigned Spectra in Shotgun Proteomics. Nat. Biotechnol 2015, 33 (7), 743–749. 10.1038/nbt.3267. [PubMed: 26076430]

(20). Na S; Bandeira N; Paek E Fast Multi-Blind Modification Search through Tandem Mass Spectrometry. Mol. Cell. Proteomics MCP 2012, 11 (4), M111.010199. 10.1074/mcp.M111.010199.

(21). Yu F; Li N; Yu W PIPI: PTM-Invariant Peptide Identification Using Coding Method. J. Proteome Res 2016, 15 (12), 4423–4435. 10.1021/acs.jproteome.6b00485. [PubMed: 27748123]

(22). Kong AT; Leprevost FV; Avtonomov DM; Mellacheruvu D; Nesvizhskii AI MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. Nat. Methods 2017, 14 (5), 513–520. 10.1038/nmeth.4256. [PubMed: 28394336]

(23). Devabhaktuni A; Lin S; Zhang L; Swaminathan K; Gonzales C; Olsson N; Pearlman S; Rawson K; Elias JE TagGraph Reveals Vast Protein Modification Landscapes from Large Tandem Mass Spectrometry Data Sets. Nat. Biotechnol 2019, 37 (4), 469–479. 10.1038/s41587-019-0067-5. [PubMed: 30936560]

(24). Yu F; Teo GC; Kong AT; Haynes SE; Avtonomov DM; Geiszler DJ; Nesvizhskii AI Identification of Modified Peptides Using Localization-Aware Open Search. Nat. Commun 2020, 11 (1), 4065. 10.1038/s41467-020-17921-y. [PubMed: 32792501]

(25). Vaudel M; Barsnes H; Berven FS; Sickmann A; Martens L SearchGUI: An Open-Source Graphical User Interface for Simultaneous OMSSA and X!Tandem Searches. Proteomics 2011, 11 (5), 996–999. 10.1002/pmic.201000595. [PubMed: 21337703]

(26). Kremer LPM; Leufken J; Oyunchimeg P; Schulze S; Fufezan C Ursgal, Universal Python Module Combining Common Bottom-Up Proteomics Tools for Large-Scale Analysis. J. Proteome Res 2016, 15 (3), 788–794. 10.1021/acs.jproteome.5b00860. [PubMed: 26709623]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(27). Jones AR; Siepen JA; Hubbard SJ; Paton NW Improving Sensitivity in Proteome Studies by Analysis of False Discovery Rates for Multiple Search Engines. Proteomics 2009, 9 (5), 1220–1229. 10.1002/pmic.200800473. [PubMed: 19253293]

(28). Kwon T; Choi H; Vogel C; Nesvizhskii AI; Marcotte EM MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines. J. Proteome Res 2011, 10 (7), 2949–2958. 10.1021/pr2002116. [PubMed: 21488652]

(29). Shteynberg D; Nesvizhskii AI; Moritz RL; Deutsch EW Combining Results of Multiple Search Engines in Proteomics. Mol. Cell. Proteomics MCP 2013, 12 (9), 2383–2393. 10.1074/mcp.R113.027797. [PubMed: 23720762]

(30). Geiszler DJ; Kong AT; Avtonomov DM; Yu F; Leprevost FV; Nesvizhski AI PTM-Shepherd: Analysis and Summarization of Post-Translational and Chemical Modifications from Open Search Results. bioRxiv 2020, 2020.07.08.192583. 10.1101/2020.07.08.192583.

(31). An Z; Zhai L; Ying W; Qian X; Gong F; Tan M; Fu Y PTMiner: Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-Translational Modification Characterization in Human Proteome. Mol. Cell. Proteomics MCP 2019, 18 (2), 391–405. 10.1074/mcp.RA118.000812. [PubMed: 30420486]

(32). Bekker-Jensen DB; Kelstrup CD; Batth TS; Larsen SC; Haldrup C; Bramsen JB; Sørensen KD; Høyer S; Ørntoft TF; Andersen CL; Nielsen ML; Olsen JV An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Syst 2017, 4 (6), 587–599.e4. 10.1016/j.cels.2017.05.009. [PubMed: 28601559]

(33). Brown CJ; Grassmyer KT; MacDonald ML; Clemmer DE; Trinidad JC Glycoproteome Analysis of Human Serum and Brain Tissue. bioRxiv 2019, 647081. 10.1101/647081.

(34). Schmidt A; Kochanowski K; Vedelaar S; Ahrné E; Volkmer B; Callipo L; Knoops K; Bauer M; Aebersold R; Heinemann M The Quantitative and Condition-Dependent Escherichia Coli Proteome. Nat. Biotechnol 2016, 34 (1), 104–110. 10.1038/nbt.3418. [PubMed: 26641532]

(35). Chambers MC; Maclean B; Burke R; Amodei D; Ruderman DL; Neumann S; Gatto L; Fischer B; Pratt B; Egertson J; Hoff K; Kessner D; Tasman N; Shulman N; Frewen B; Baker TA; Brusniak M-Y; Paulse C; Creasy D; Flashner L; Kani K; Moulding C; Seymour SL; Nuwaysir LM; 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 EW; Moritz RL; Katz JE; Agus DB; MacCoss M; Tabb DL; Mallick P A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. Nat. Biotechnol 2012, 30 (10), 918–920. 10.1038/nbt.2377. [PubMed: 23051804]

(36). Schulze S; Adams Z; Cerletti M; De Castro R; Ferreira-Cerca S; Fufezan C; Giménez MI; Hippler M; Jevtic Z; Knüppel R; Legerme G; Lenz C; Marchfelder A; Maupin-Furlow J; Paggi RA; Pfeiffer F; Poetsch A; Urlaub H; Pohlschroder M The Archaeal Proteome Project Advances Knowledge about Archaeal Cell Biology through Comprehensive Proteomics. Nat. Commun 2020, 11 (1), 3145. 10.1038/s41467-020-16784-7. [PubMed: 32561711]

(37). Kösters M; Leufken J; Schulze S; Sugimoto K; Klein J; Zahedi RP; Hippler M; Leidel SA; Fufezan C PymzML v2.0: Introducing a Highly Compressed and Seekable Gzip Format. Bioinformatics 2018, 34 (14), 2513–2514. 10.1093/bioinformatics/bty046. [PubMed: 29394323]

(38). Craig R; Beavis RC TANDEM: Matching Proteins with Tandem Mass Spectra. Bioinforma. Oxf. Engl 2004, 20 (9), 1466–1467. 10.1093/bioinformatics/bth092.

(39). Kim S; Mischerikow N; Bandeira N; Navarro JD; Wich L; Mohammed S; Heck AJR; Pevzner PA The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search. Mol. Cell. Proteomics MCP 2010, 9 (12), 2840–2852. 10.1074/mcp.M110.003731. [PubMed: 20829449]

(40). Käll L; Canterbury JD; Weston J; Noble WS; MacCoss MJ Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. Nat. Methods 2007, 4 (11), 923–925. 10.1038/nmeth1113. [PubMed: 17952086]

(41). Liu M-Q; Zeng W-F; Fang P; Cao W-Q; Liu C; Yan G-Q; Zhang Y; Peng C; Wu J-Q; Zhang X-J; Tu H-J; Chi H; Sun R-X; Cao Y; Dong M-Q; Jiang B-Y; Huang J-M; Shen H-L; Wong CCL; He S-M; Yang P-Y PGlyco 2.0 Enables Precision N-Glycoproteomics with Comprehensive Quality Control and One-Step Mass Spectrometry for Intact Glycopeptide Identification. Nat. Commun 2017, 8 (1), 438. 10.1038/s41467-017-00535-2. [PubMed: 28874712]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(42). Eichler J; Arbiv A; Cohen-Rosenzweig C; Kaminski L; Kandiba L; Konrad Z N-Glycosylation in Haloferax Volcanii: Adjusting the Sweetness. Front. Microbiol 2013, 4. 10.3389/fmicb.2013.00403.

(43). Khatri K; Klein JA; Zaia J Use of an Informed Search Space Maximizes Confidence of Site-Specific Assignment of Glycoprotein Glycosylation. Anal. Bioanal. Chem 2016. 10.1007/s00216-016-9970-5.

(44). Bekker-Jensen DB; Kelstrup CD; Batth TS; Larsen SC; Haldrup C; Bramsen JB; Sørensen KD; Høyer S; Ørntoft TF; Andersen CL; Nielsen ML; Olsen JV An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Syst 2017, 4 (6), 587–599.e4. 10.1016/j.cels.2017.05.009. [PubMed: 28601559]

(45). Avtonomov DM; Kong A; Nesvizhskii AI DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. J. Proteome Res 2019, 18 (2), 715–720. 10.1021/acs.jproteome.8b00728. [PubMed: 30523686]

(46). Bubis JA; Levitsky LI; Ivanov MV; Gorshkov MV Validation of Peptide Identification Results in Proteomics Using Amino Acid Counting. Proteomics 2018, 18 (23), e1800117. 10.1002/pmic.201800117. [PubMed: 30307114]

(47). Kertesz-Farkas A; Keich U; Noble WS Tandem Mass Spectrum Identification via Cascaded Search. J. Proteome Res 2015, 14 (8), 3027–3038. 10.1021/pr501173s. [PubMed: 26084232]

(48). Izaham ARA; Scott NE Open Database Searching Enables the Identification and Comparison of Bacterial Glycoproteomes without Defining Glycan Compositions Prior to Searching. Mol. Cell. Proteomics 2020, 19 (9), 1561–1574. 10.1074/mcp.TIR120.002100. [PubMed: 32576591]

(49). Polasky DA; Yu F; Teo GC; Nesvizhskii AI Fast and Comprehensive N- and O-Glycoproteomics Analysis with MSFragger-Glyco. bioRxiv 2020, 2020.05.18.102665. 10.1101/2020.05.18.102665.

(50). Lu L; Riley NM; Shortreed MR; Bertozzi CR; Smith LM O-Pair Search with MetaMorpheus for O-Glycopeptide Characterization. bioRxiv 2020, 2020.05.18.102327. 10.1101/2020.05.18.102327.

(51). Medzihradszky KF; Kaasik K; Chalkley RJ Tissue-Specific Glycosylation at the Glycopeptide Level. Mol. Cell. Proteomics MCP 2015, 14 (8), 2103–2110. 10.1074/mcp.M115.050393. [PubMed: 25995273]

(52). Solntsev SK; Shortreed MR; Frey BL; Smith LM Enhanced Global Post-Translational Modification Discovery with MetaMorpheus. J. Proteome Res 2018, 17 (5), 1844–1851. 10.1021/acs.jproteome.7b00873. [PubMed: 29578715]

(53). Chang H-Y; Kong AT; da Veiga Leprevost F; Avtonomov DM; Haynes SE; Nesvizhskii AI Crystal-C: A Computational Tool for Refinement of Open Search Results. J. Proteome Res 2020, 19 (6), 2511–2515. 10.1021/acs.jproteome.0c00119. [PubMed: 32338005]

(54). Polasky DA; Yu F; Teo GC; Nesvizhskii AI Fast and Comprehensive N - and O - Glycoproteomics Analysis with MSFragger-Glyco. Nat. Methods 2020, 1–8. 10.1038/s41592-020-0967-9. [PubMed: 31907477]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

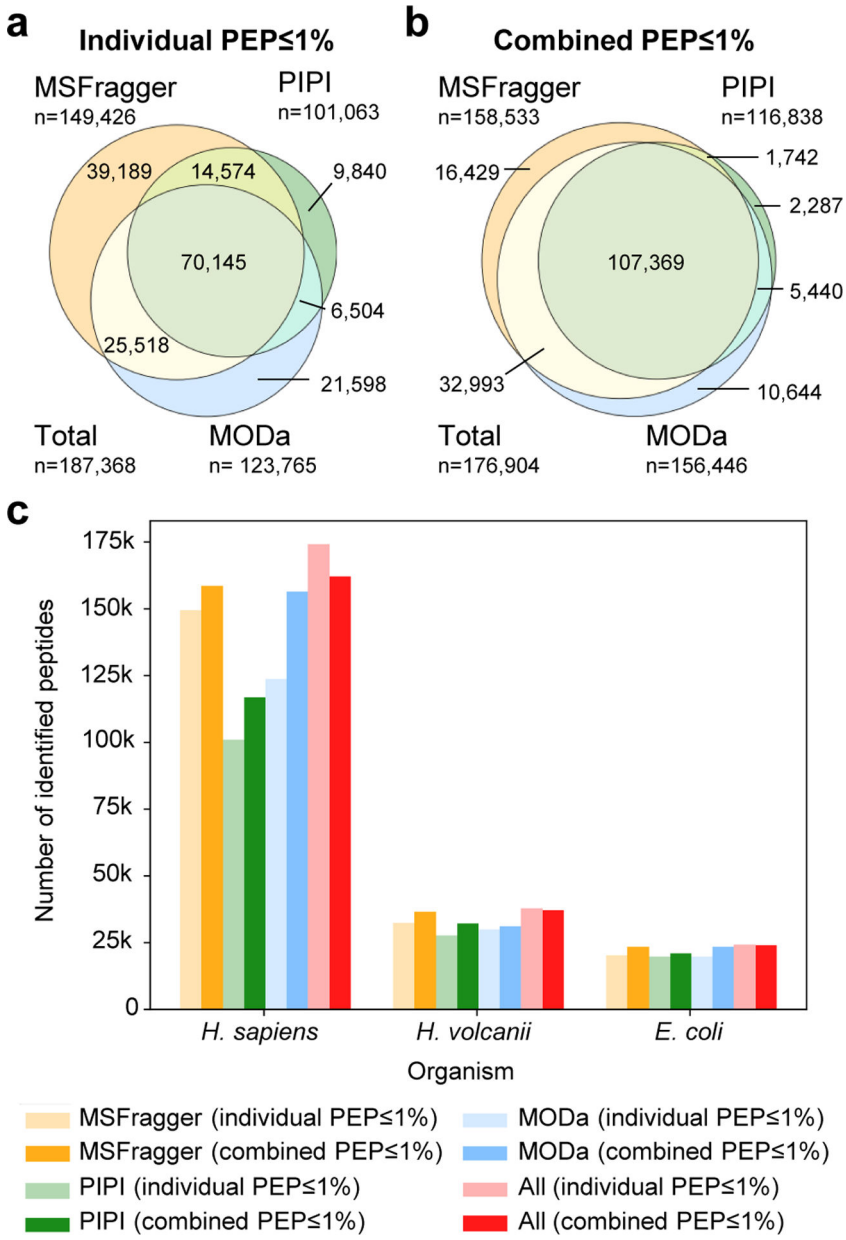



**Figure 1. The combination of results from different OMS engines lead to an increase in identified peptide sequences.**

Venn diagrams comparing peptide sequences identified by the OMS engines MSFragger (orange), MODa (blue) and PIPI (green) are shown, considering only PSMs filtered by 1% PEP for each engine separately (**a**), and representing results from a combined PEP approach with PSMs filtered by 1% combined PEP (**b**). Both Venn diagrams show results for the human dataset. Comparable results for *E. coli* and *H. volcanii* are presented in Fig. S1. (**c**), Across three studied datasets, the number of identified peptide sequences based on PSMs filtered by 1% PEP for each engine (light colors) is compared to those found in PSMs filtered by 1% combined PEP (dark colors). Results are presented for each individual engine as well as the merged results of all engines (red). It should be noted that for merged results of PSMs filtered by 1% individual engine PEP, the overall PEP may be as high as 3%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results were sanitized (using Ursgal, see Methods for details), accepting only the best scoring PSM per spectrum. For (a) and (b) this sanitizing step was performed at the single engine level, to represent peptide sequences identified by each individual engine. For merged results of all engines within (c) however, sanitizing was performed over results from all engines, accepting only the best scoring PSM per spectrum for conflicting matches between different engines.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

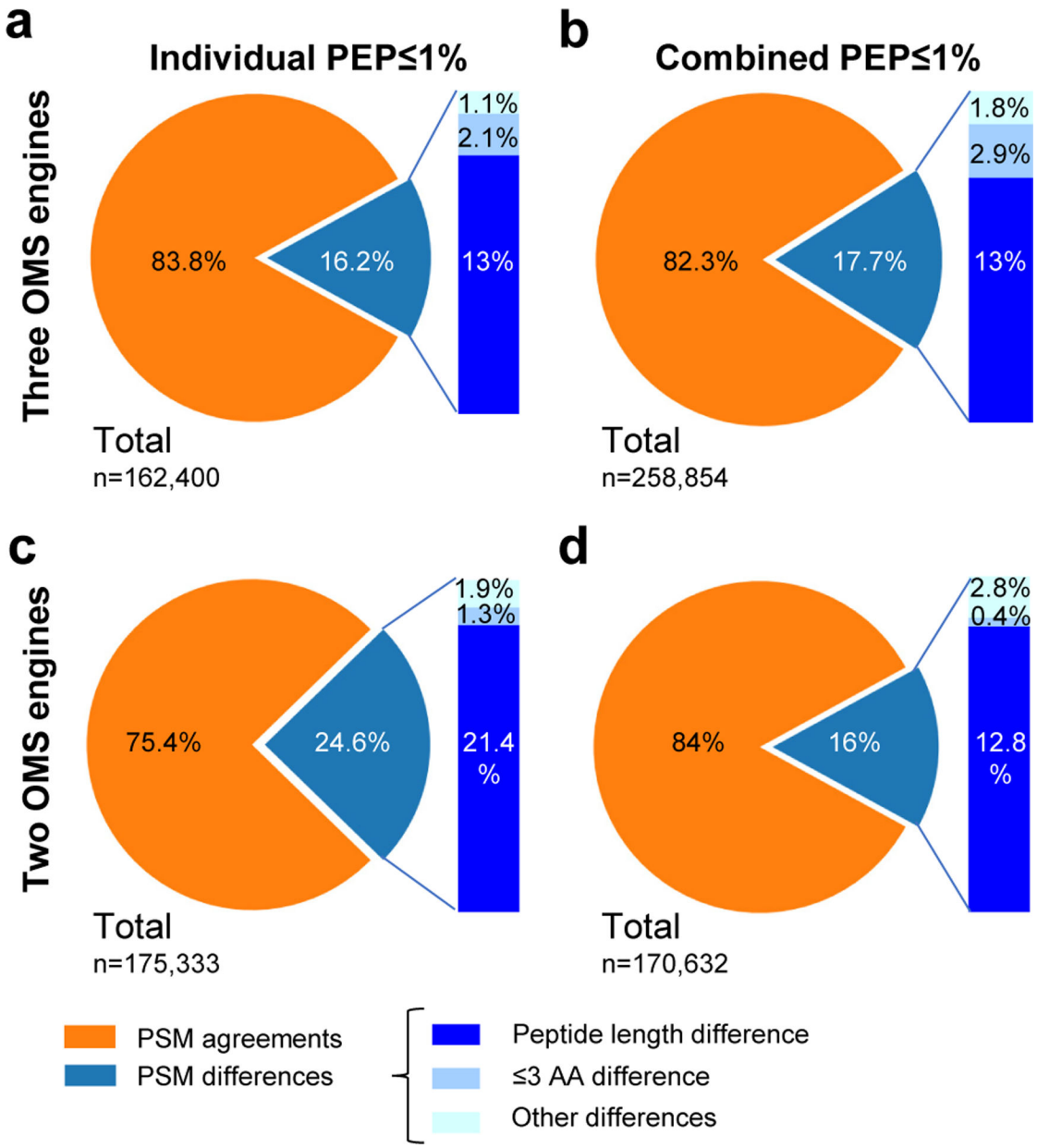



**Figure 2. Differences in PSM assignments between different OMS engines were mainly driven by differences in peptide length and reduced by the combined PEP approach.**

Spectra for which PSMs were identified by all three engines (**a**, **b**) as well as spectra for which PSMs were identified by two engines (**c**, **d**) were analyzed for agreements (orange) and differences (blue), i.e. whether at least one engine identified a different peptide sequence than the other(s). PSMs were filtered by 1% individual engine PEP in (a) and (c) in contrast to a filtering by 1% combined PEP in (b) and (d). Spectra with PSM differences were sorted into three categories: (i) peptide sequences with differing lengths, for which the shorter peptide is a substrings of the longer one (dark blue); (ii) peptides with one to three differing amino acids (medium blue); and (iii) peptides differing in any other way (light blue, considered clear disagreements). The results here represent the human dataset. Similar results for *E. coli* and *H. volcanii* are shown in Fig. S3.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

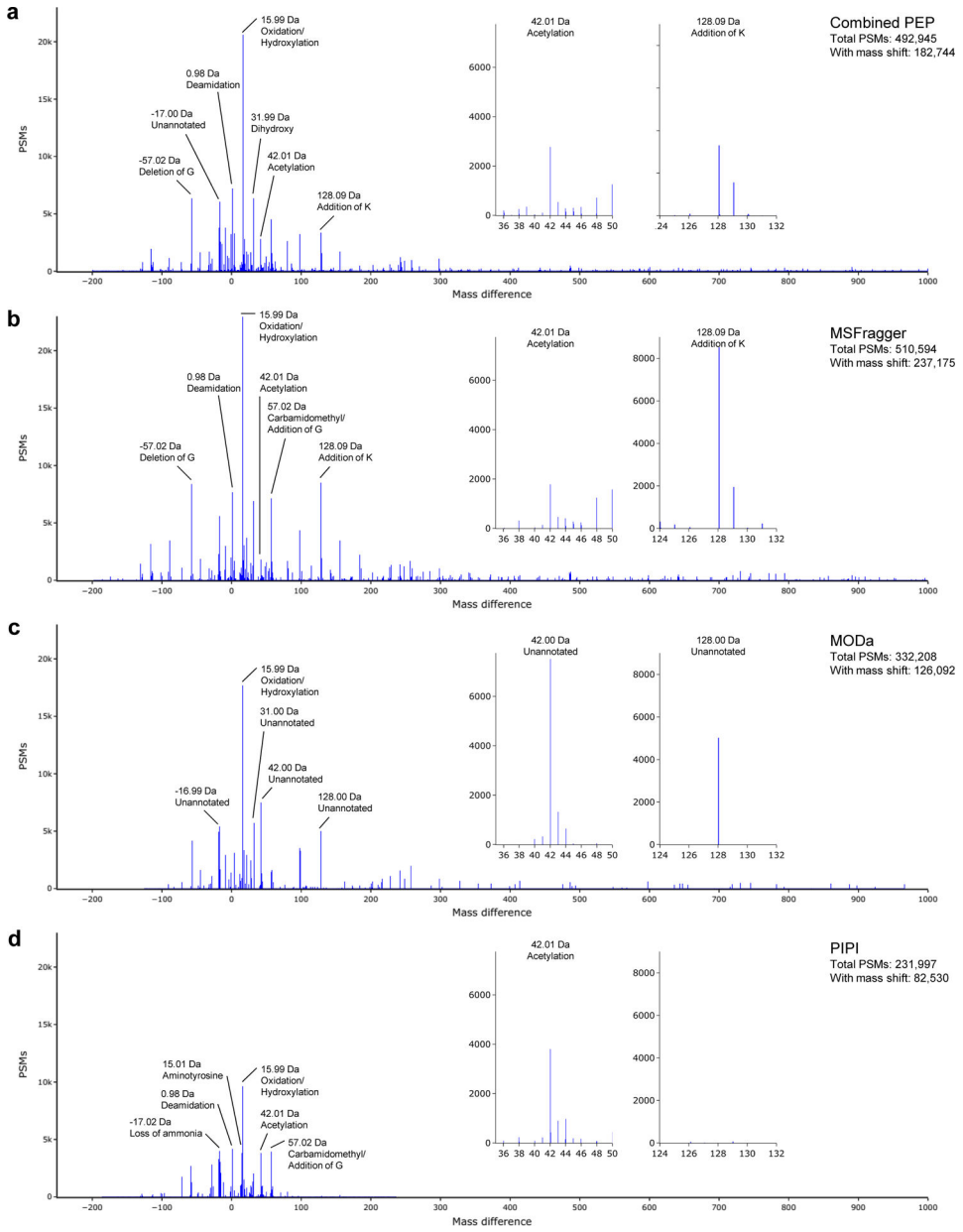



**Figure 3. The unified post-processing of OMS results assisted in interpreting varied mass difference profiles of different OMS engines.**

Mass differences from combined results of all employed OMS engines (**a**), filtered by 1% combined PEP, as well as mass differences from results of MSFragger (**b**), MODa (**c**), and PIPI (**d**), filtered by 1% individual PEP, were post-processed using PTM-Shepherd. Results are presented as histograms giving the number of PSMs for each 0.05 Da mass difference bin, omitting PSMs with no mass difference, or mass differences corresponding to isotopic peak selection. Results were sanitized on the individual engine level (b-c) or for the combination of all engines (a) before post-processing with PTM-Shepherd. Annotations for the most prominent mass differences are given. Insets represent enlarged areas of the mass differences corresponding to acetylation (left) and addition of K (right), highlighting differences between the engines and the combined PEP approach. The number of all

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

identified PSMs ("Total PSMs") is given for all subfigures as well as the number of PSMs that contain mass shifts shown in the profiles ("With mass shift"). It should be noted that PSMs containing multiple mass differences are counted multiple times. While results for the *H. sapiens* dataset are shown here, mass difference profiles for the *E. coli* and *H. volcanii* datasets can be found in Fig. S5 and S6, respectively.

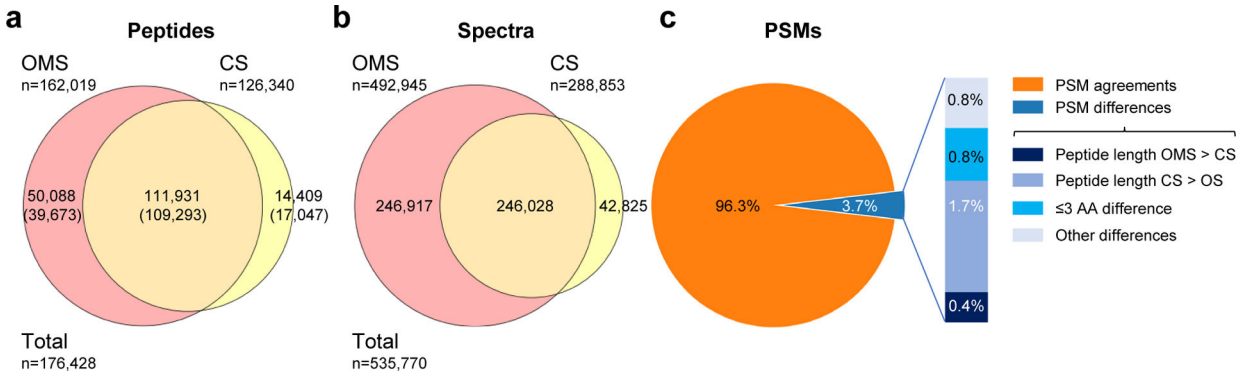



**Figure 4. OMS, in comparison to CS, increased the number of identified peptide sequences and spectra, including the vast majority of identifications by CS.**

Combined results from three OMS engines (red; MSFragger, MODa and PIPI) are compared to combined results from three CS engines (yellow; MSFragger, MS-GF+ and X!Tandem) for identified peptide sequences (**a**) and spectra (**b**). Results were filtered by 1% combined PEP and sanitized for each approach separately. Of the 246,028 commonly identified spectra, the percentage of PSM agreements (orange) and differences (blue) between the OMS and CS approach is given (**c**). Spectra with PSM differences are put in four categories: (i) spectra for which OMS engines matched a peptide sequences that is a substring of the one matched by CS engines (peptide length CS > OMS); (ii) spectra for which CS engines matched a peptide sequences that is a substring of the one matched by OMS engines (peptide length OMS > CS); (iii) spectra for which the peptide sequences matched by OMS and CS engines differed by one to three amino acids; and (iv) spectra with peptide sequences differing between OMS and CS in any other way (considered clear disagreements). These results correspond to the *H. sapiens* dataset. Fig. S7 shows similar outcomes for *E. coli* and *H. volcanii* datasets.



Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

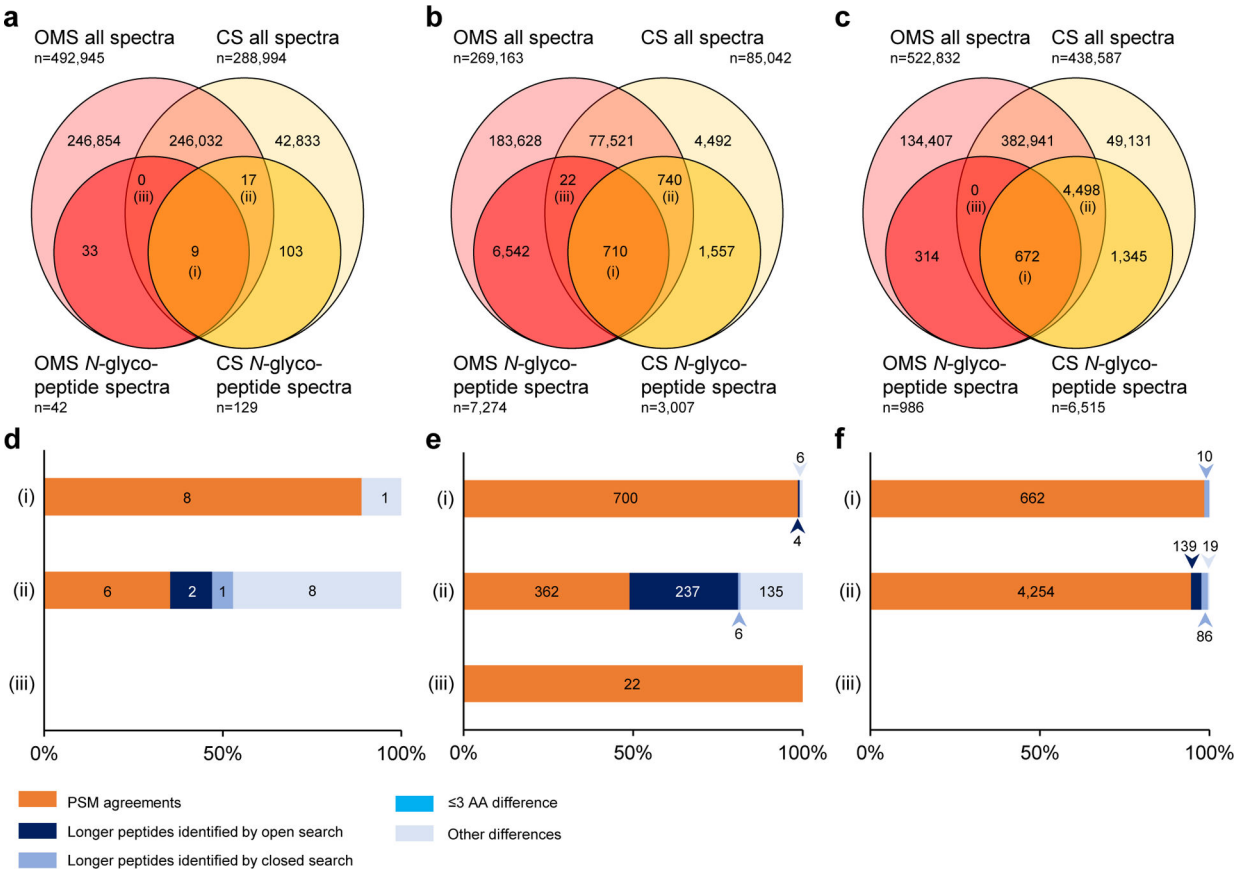



**Figure 5. The identification of complex N-glycopeptides remains challenging for OMS engines.**

Venn diagrams are shown comparing all spectra identified by the combined OMS approach (light red), with a subset of spectra matched to peptides containing mass shifts that were mapped to known *N*-glycans (dark red), and all spectra identified by the combined CS approach (light yellow), with a subset of spectra matched to *N*-glycopeptides (dark yellow). These comparisons are shown for (**a**), the *H. sapiens* in-depth proteomics dataset (PXD004452), (**b**) the *H. sapiens* glycoproteomic dataset (PXD013715), and (**c**), the *H. volcanii* dataset (PXD021874). Results were filtered by 1% combined PEP and sanitized for each approach separately. For each subfigure, three groups of spectra are analyzed in more detail: (i) spectra with *N*-glycopeptides identified in both approaches, CS and OMS; (ii) spectra matched to *N*-glycopeptides by CS engines but for which OMS engines identified peptides with mass shifts that could not be mapped to known *N*-glycans; and (iii) spectra matched by OMS engines to peptides with mass shifts that are annotated as *N*-glycans while the CS approach did not match them to *N*-glycopeptides. (**d**), (**e**), and (**f**) present PSM agreements (orange) and differences (blue) in each of these three categories for the datasets corresponding to (a), (b), and (c), respectively. The following types of differences are considered: peptide sequences with differing lengths, for which either OMS or CS identified the shorter peptide that is a substrings of the longer one; peptides with one to three differing amino acids; and peptides differing in any other way (considered clear disagreements). It should be noted that *H. volcanii* glycopeptides harbor linear *N*-glycans with up to five monosaccharides, i.e. less complex *N*-glycans than *H. sapiens'*. Furthermore, results from

CSs can contain multiple glycans per peptide, while combinations of glycans have not been included in the matching of OMS results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 1.**

OMS parameters used for all analyzed datasets.

| Dataset | Maximum modification size (Da) | Minimum modification size (Da) | Precursor mass tolerance (ppm)* | Fragment mass tolerance (ppm) |
|---|---|---|---|---|
| *H. sapiens* (PXD004452) | +4000 | −200 | ±5 | ±20 |
| *H. sapiens* (PXD013715) | +4000 | −200 | ±10 | ±10 |
| *H. volcanii* (PXD021874) | +2000 | −2000 | ±10 | ±10 |
| *E. coli* (PXD000498) | +2000 | −2000 | ±5 | ±20 |

*For OMSs using MSFragger, the maximum/minimum modification size is technically given as precursor mass tolerance while the final tolerance for PSMs is given by the precursor true mass tolerance.