Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions

Ana S C Silva [1,2], Robbin Bouwmeester [1,2], Lennart Martens [1,2§],
Sven Degroeve[1,2]

[1] VIB-UGent Center for Medical Biotechnology, Ghent, Belgium

[2] Department of Biomolecular Medicine, Faculty of Medicine, Ghent, Belgium

[§]To whom correspondence should be addressed:
lennart.martens@ugent.vib.be

Bioinformatics, 11 May 2019
https://doi.org/10.1093/bioinformatics/btz383

The use of post-processing tools to maximize the information gained from a proteomics search engine is widely accepted and used by the community, with the most notable example being Percolator - a semi-supervised machine learning model which learns a new scoring function for a given dataset. The usage of such tools is however bound to the search engine's scoring scheme, which doesn't always make full use of the intensity information present in a spectrum. We aim to show how this tool can be applied in such a way that maximizes the use of spectrum intensity information by leveraging another machine learning-based tool, MS2PIP. MS2PIP predicts fragment ion peak intensities. We show how comparing these intensities to annotated experimental spectra by calculating direct similarity metrics provides enough information for a tool such as Percolator to accurately separate two classes of PSMs. This approach allows using more information out of the data (compared to simpler intensity based metrics, like peak counting or explained intensities summing) while maintaining control of statistics such as the false discovery rate.

1 INTRODUCTION

Proteomics is a field that relies heavily on mass spectrometry for the identification of proteins in a sample (Aebersold and Mann, 2003). The identification process of the acquired fragmentation mass spectra is carried out using bioinformatics tools called search engines that match experimentally obtained mass spectra to peptide sequences (Verheggen *et al.*, 2017).

Sequence database search engines, who are by far the most popular kind of search engine, assign sequences to spectra by generating theoretical spectra for each potential sequence, matching it to the experimental spectra and attributing a score to each match (Verheggen *et al.*, 2017). These theoretical spectra are typically very simple: e.g. SEQUEST (Eng *et al.*, 1994) creates them by assigning an arbitrary magnitude of 50 for b- and y-ion fragment peaks, of 25 for ions with m/z equal to ±1 u from the b- and y-ion fragments, and of 10 for ions associated with neutral losses of water or ammonia. These different intensities aim to represent the relative abundances of the different ion types. Even though some distinction is thus hinted at, this approach is far from representing the real differences in intensity seen experimentally.

A score is then calculated based on the match between such a theoretical spectrum and an experimental spectrum. This peptide-to-spectrum match (PSM) score is intended to allow the discrimination of truthful matches from random matches. This is done through scoring functions that can be based on the number of matched peaks, or on the sum of matched experimental peak intensity. These scoring functions typically include a statistical model that attempts to describe the probability of obtaining a match at this score by chance (Yilmaz *et al.*, 2017). The output of a search engine consists of PSMs with scores to indicate their reliability against a random match model, which typically take the form of an e-value or similar statistical metric.

Although this approach delivers reliability metrics for individual PSMs, it does not address the overall number of incorrect PSMs that can be expected to occur when many PSMs are reported. To address this issue, a correction for multiple testing needs to be performed, and this is typically implemented as a false discovery rate (FDR) control that is calculated across the top-scoring PSMs for each spectrum, by comparing the experimental PSM's score distribution to a null distribution of scores (Benjamini and Hochberg, 1995; Eriksson *et al.*, 2000; Nesvizhskii, 2010; Sticker *et al.*, 2017).

In proteomics, this null distribution is not based on statistical models, but derived from empirical data. This is achieved by searching a compound target-decoy database, in which the target database contains the protein sequences of biological interest, whereas the decoy database contains non-sensical sequences that are designed to accurately represent false positive identification results.

In order to improve the amount of spectra identified at a fixed FDR, researchers frequently couple search engines to specially developed post-processing tools. These tools employ machine learning algorithms to separate true from false PSMs by exploiting all the information available of the PSM in the form of feature vectors. This information includes PSM scores computed by the search engine, delta scores, peptide mass, charge state, precursor mass error, fragmentation mass errors etc. Post-processing tools have been available from as early as 2002 (Keller *et al.*, 2002), and have since gained increasing popularity in the normal proteomics bioinformatics workflow.

The earliest attempts at tackling the PSM identification problem from a data-driven perspective framed it as a binary classification problem, where the positive class is represented by confident target PSMs, whereas the negative class is represented by decoy PSMs. One such attempt was Anderson *et al.*'s

(2003), who trained a support vector machine (SVM) on PSMs obtained from SEQUEST (Eng *et al.*, 1994). The SVM was trained on thirteen PSM features (of which nine were obtained directly from the search engine's score calculations), with the positive class of PSMs selected by expert criteria, whereas the negative class corresponded to PSMs associated with incorrect proteins.

Percolator (Käll *et al.*, 2007) perfects this approach. Although similar in principle, it introduces some crucial differences that have made it into the most popular post-processing tool in the proteomics community (The *et al.*, 2016). The first of these differences was the re-framing of the problem as a semi-supervised task: after a search engine run, the results include three rather than two classes: 'negative', represented by the decoy hits; 'positive', represented by the target hits above the FDR threshold; and 'unknown', represented by the target hits which fall under the FDR threshold. An SVM is trained on the positive and negative classes in order to learn a (scoring) function that best separates these two. The second crucial difference is that this new scoring function is not meant to generalize to other datasets, but only to rescore the PSMs in the current dataset. This includes the PSMs in the 'unknown' class, and can result in members of this class being re-assigned to the positive class, recovering more PSMs at the same FDR cutoff.

Here, we replace the search engine features with spectral comparison features, with the goal of overcoming post-processing tools' dependency on the search engine's score calculations and instead maximizing the use of intensity information present in a spectrum. To this end we make use of fragmentation spectrum intensity predictions by MS2PIP (Degroeve and Martens, 2013) to calculate an extensive set of features that provide additional discriminative and complementary information for PSM post-processing. By comparing the intensity information in these predicted spectra with those recorded in the experimental spectra, our tool can be used to sensitively identify truthful PSMs while maintaining controlled FDR. This allows going beyond explained peak counting or summing explained peak intensities: fragment ion peaks with predicted low intensities should have a positive contribution to the PSM score if that peak was indeed observed with low intensity, something that is not the case for current search engine PSM scoring functions.

We moreover show that these features are sufficient to render this approach effectively independent of the original search engine score and related features, thus making it compatible with any desired search engine, or with the combined output of multiple search engines, as for instance obtained by PeptideShaker (Vaudel *et al.*, 2015).

2 MATERIALS AND METHODS

2.1 MS2PIP

MS2PIP is a data-driven tool that predicts fragment ion intensities given a (modified) peptide sequence and charge. MS2PIP is comprised of different models for different fragmentation types; definite models for singly charged fragment ions exist for high–energy collisional dissociation (HCD) and collision–induced dissociation (CID) fragmentation, while models for electron–transfer and higher–energy collision dissociation (EThcD) fragmentation, doubly charged fragment ions in HCD, tandem mass tag (TMT) and isobaric tags for relative and absolute quantitation (iTRAQ) labeled data and triple time of flight (TTOF) mass analyzers are also available in beta. The performance of each of these models can be found in CompOmics (2019b) and Gabriels *et al.* (2019). Since its initial publication, MS2PIP has gone through several updates, and is currently available as a web-service (Gabriels *et al.*, 2019) and as a locally installable package (CompOmics, 2019a).

The input for MS2PIP is a PEPREC (PEPtide RECord; CompOmics, 2019) file. This file contains peptide sequences, post-translational modifications (PTMs) and charges, each associated to a unique spectrum identifier. MS2PIP reads this file, and predicts the intensity for each theoretical fragment ion. If MS2PIP is provided with experimental spectra through an mgf file (mascot generic format; Perkins *et al.*, 1999), and there is a correspondence between spectrum identifiers in the PEPREC file and the 'TITLE' field of the spectrum file, it outputs both the experimental intensity (obtained from the mgf file) and the predicted intensity for each fragment ion.

2.2 Feature extraction and selection

A machine learning method's performance is strongly tied to the features used to describe the problem it will try to solve. Several metrics can be used to compare the experimental and empirical spectrum pairs that we have for each PSM, and a large number of features can therefore be devised.

MS2PIP-predicted spectra are *log*2 transformed total-ion-current (TIC) normalized intensities for the theoretical fragment ions. Given the nature of the *log*2 transformation, low intensity peaks are weighed relatively more heavily by spectrum comparison metrics. Spectral correlation features are therefore calculated twice: once on the *log*2 transformed spectra, and once on the TIC-normalized spectra in the original linear scale. This allows different similarities and differences to be emphasized by the two distinct intensity scales. Moreover, these correlation calculations are not only performed for the entire spectrum but also for the b- and y-ion series separately. The resulting set of features is given in Supplementary Table S1. Note that this set of features will include significant redundancy. Despite the model used being an SVM, which includes regularization strategies that can handle issues that can stem from this redundancy, we choose to select features based on their correlations. By eliminating features that correlate to other features with a Pearson correlation coefficient of 0.9 or greater, we reduce the feature set to a total of 44 features (around 60% of its initial size). The selected features can be found in Supplementary Material S1.

2.3 Datasets and processing

To compare the different feature sets we constructed entrapment protein databases as described in Vaudel *et al.* (2012). The proposed procedure consists of searching a *Pyrococcus furiosus* (Pfu) sample, as obtained from the PRIDE archive (Vaudel, 2014; Vizcaíno *et al.*, 2016), against a target protein database that consists of the Pfu proteome (reviewed and unreviewed sequences, for a total of 5087 proteins) and a database of entrapment proteins from an eukaryotic organism. A decoy database is then constructed by reversing this concatenated target database. Every true PSM should be a match against the Pfu proteome as the evolutionary distance between Pfu and eukaryotes was shown to be sufficiently large to avoid peptide overlap. Every random (false) PSM should match an entrapment or a decoy peptide with equal probability (given a large enough entrapment database we can neglect the contribution of the Pfu proteins). Overfitting on the target database can now be detected by computing the ratio of identified PSMs at 1% FDR (i.e. 1% matches against decoy peptides) that match an entrapment peptide. This should be equal or smaller than 1%.

Two target databases were constructed by collecting all Pfu proteins and concatenating it to all *Homo sapiens* sequences (reviewed, 22 449 proteins) and for the other database all eukaryota sequences (reviewed, 188 993 proteins) from The UniProt Consortium (2017).

The spectra are searched against these databases using MS-GF+ (Kim and Pevzner, 2014). More details on the search settings can be found in Supplementary Materials S2 and S3. The search results (i.e. all rank 1 PSMs) were converted into a Percolator input file (pin, from Käll *et al.*, 2007), which contains Percolator features for each PSM. These features can be split into two main groups: features based on the peptide sequence (such as amino-acid frequency, peptide length and precursor charge), and features based on the search engine's score (such as, in this case, the scoring metrics calculated by MS-GF+). Percolator is executed on these searches, so that a 'baseline improvement' can be considered and compared against our proposed approach.

From the list of all PSMs obtained by MS-GF+, an input file for MS2PIP (a PEPREC file) was created. This file was ran through MS2PIP, along with the mgf file containing the original spectra, producing a file with pairs of spectra (predicted and experimental) for each input PSM sequence. From these pairs, a matrix of features was calculated, where each row corresponds to a PSM and each column to a feature metric for that PSM from the list in Supplementary Material S1. This matrix was combined with the pin file so that both the default Percolator features and the spectral comparison features are available for each PSM. Then Percolator was run with this compound matrix as input. In addition, Percolator was also run using only the MS2PIP features, which thus notably excludes all MS-GF+ scoring information.

Besides these entrapment experiments, which function as a validation of our proposed feature sets, we show the results obtained when applying our procedure in two larger and more noisy datasets. These were obtained from PRIDE (Chick, 2015; Kim *et al.*, 2014); the first contains a human embryonic kidney HEK) cell line, and the second is a selection from the draft of the human proteome—specifically, all datasets from an adult adrenal gland. Search settings used were based on the information included in the PRIDE repository, and can be found in Supplementary Materials S4 and S5, respectively.

The code and data necessary to reproduce these results are available on GitHub (CompOmics, 2019), along with installation and usage instructions.

3 RESULTS

3.1 Entrapment experiments

The discriminative power of the spectral intensity prediction features was evaluated by comparing the results of the MS-GF+ search engine with three different post-processing identification strategies, each using a different set of features: the default Percolator features (percolator-set), only MS2PIP-derived intensity correlation features (ms2pip-set), and the Percolator and MS2PIP-derived features combined (both-set). To verify the feature sets' robustness, two entrapment experiments were performed as described in the previous section. These results are described in this section.

The first entrapment experiment consisted of the addition of all *H.sapiens* sequences to the database the Pfu dataset is searched against. Figure 1 shows the outcome of this experiment. Although the bars' lengths represent the number of identifications, the statistics inside each bar show the number of identified entrapment matches and the ratio of identified entrapment matches over the total number of identifications. We observe that, of the three re-scoring feature sets, the both-set computes the most identifications with the lowest ratio of entrapment PSMs. However, compared with MS-GF+ we do see an increase of entrapment PSMs for all three feature sets. Note that the entrapment ratio is well below 1% for all approaches which is expected as in this case the target database contains relatively few entrapment proteins compared with the Pfu proteome.

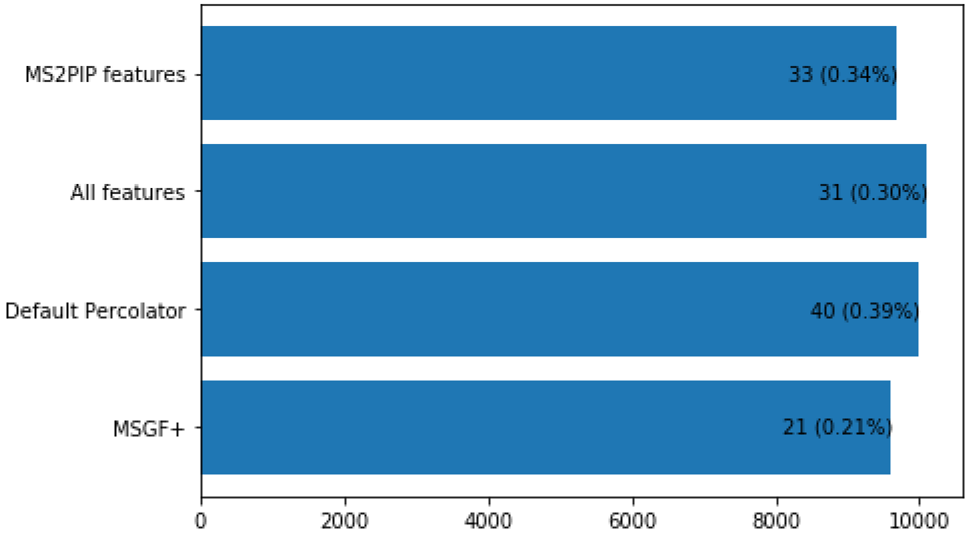

Figure 1: Number of identified PSMs (1% FDR) for each feature set (and for the search engine by itself), with amount and percentage of entrapment identifications written into each bar, for the *Homo sapiens* entrapment experiment.

In all four cases (i.e. the three features sets and the search engine), the amount of entrapment PSMs in the results is well under the FDR estimated with the decoy sequences; this indicates that all the feature sets used to score PSMs are able to provide a good estimate of the null distribution with the decoy sequences—otherwise we would expect to see more entrapment sequences in the FDR-filtered results.

To provide additional insight into the performance of the feature sets we plotted the number of identified PSMs at different rates of the FDR in Figure 2. We observe that the differences between the feature sets become more significant when we decrease the allowed FDR, with the both-set remaining 84% of the identifications at 0.01% FDR compared with 85% at 1% FDR.

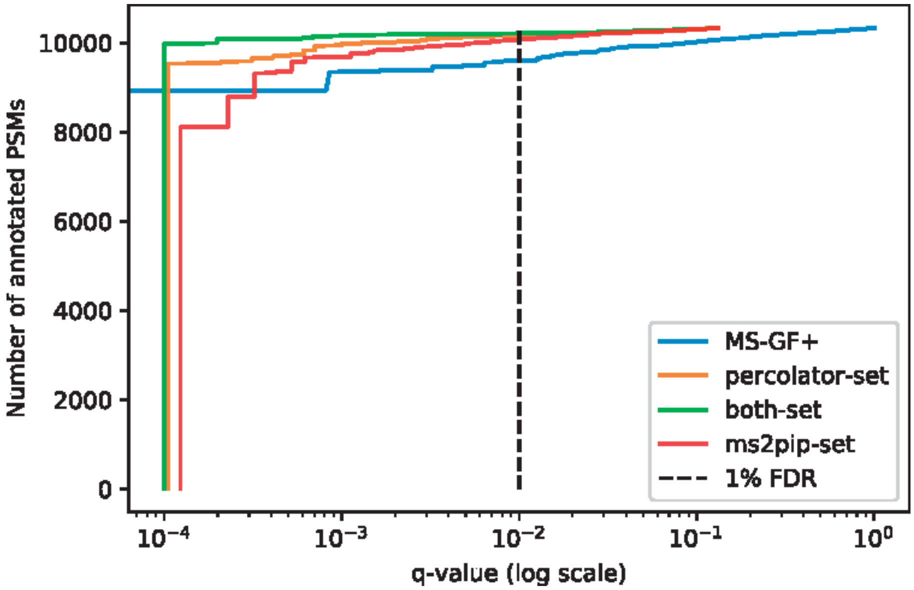

Figure 2: Number of identified PSMs at different levels of FDR for each feature set and for the search engine by itself for the H.sapiens entrapment experiment. The x-axis shows the number of PSMs, whereas the y-axis shows the q-values (in log-scale)

To obtain further insight into the differences between the feature sets we looked at the differences between observed and predicted chromatographic retention times (RTs) for uniquely identified PSMs.

ELUDE (Moruz *et al.*, 2010) was trained on 90% of the PSMs identified by all four methods (10% was used as test set) and applied to the PSMs uniquely identified by each method to compute the RT predictions for the PSMs. This process is further detailed in Supplementary Material S6.

The median absolute prediction error is 113 s for the 10% test set. For the ms2pip-set this is 111 s while for the percolator-set this increases by 55% to 172 s. For MS-GF+ and both-set the error is very similar at 148 and 149 s, respectively. These observations strongly indicate that the features computed from the MS2PIP predictions allow for a reduction of identified PSMs with large RT error, while increasing the number of identified PSMs with small RT error. Note that none of the methods exploit the PSM RT at any point.

Proceeding this experiment, a much larger entrapment database is used (as described in Section 2.3). Large databases are known to give rise to issues when using classic spectrum identification techniques (Muth *et al.*, 2015), so a decrease in general performance is expected. However, the additional spectrum comparison-based features provide an additional layer of information that can be used to 'rescue' identifications in these situations. The plot in Figure 3 shows the same comparison as Figure 1 but for this second experiment.

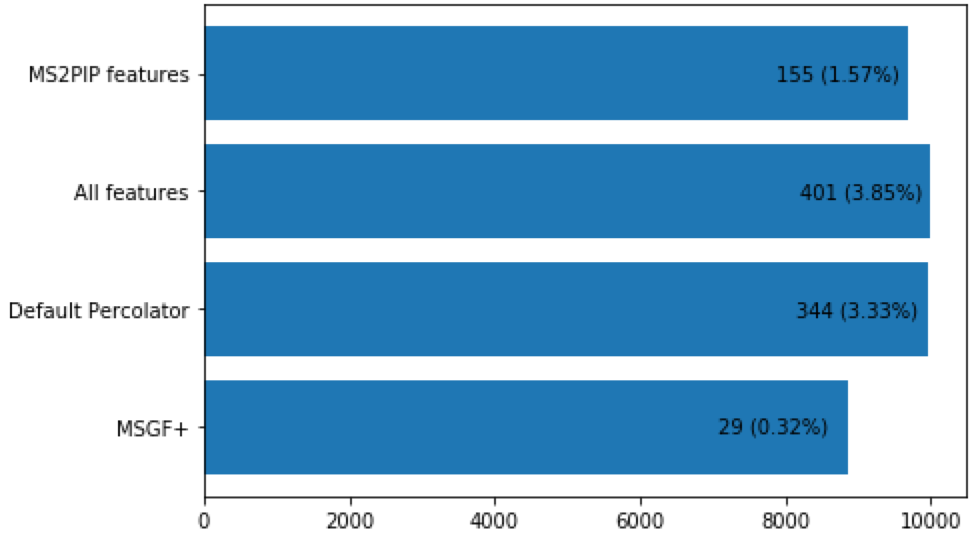

Figure 3: Number of identified PSMs (1% FDR) for each feature set (and for the search engine by itself), with amount and percentage of entrapment identifications written into each bar, for the eukaryota entrapment experiment

Immediately it is noticed that the amount of PSMs obtained by MS-GF+ is much lower than when a smaller entrapment database was used; however, the search engine manages to control the FDR, maintaining the amount of entrapment PSMs well below the estimated 1%. Although the three post-processors manage to increase the number of PSMs to the same value range as what was obtained in the previous experiment, they allow many more entrapment PSMs through, showing the effect of the database size increment in such approaches. However, using the MS2PIP-based features alone it is possible to obtain a balance, as we can observe an amount of PSMs on par as the previous experiment, and an amount of entrapment PSMs only marginally above the estimated FDR. This increase in number of PSMs can also be observed for other levels of FDR, as can be seen in Figure 4.

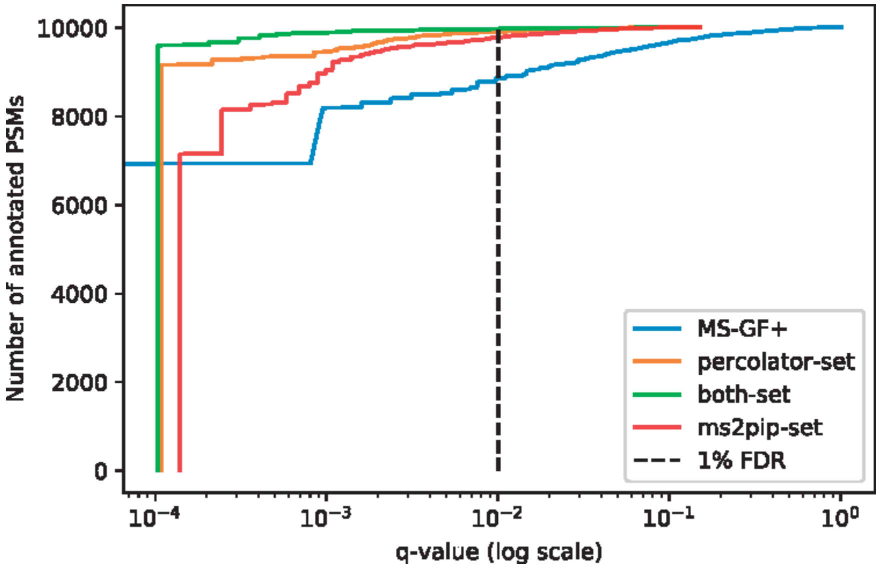

Figure 4: Number of identified PSMs at different levels of FDR for each feature set and for the search engine by itself for the eukaryota entrapment experiment. The x-axis shows the number of PSMs, whereas the y-axis shows the q-values (in log-scale)

We again trained ELUDE RT prediction models on all the PSMs that the four methods agreed on, for a total of 8733 peptides. Details on the trained model can be found in Supplementary Material S7. For this experiment, the median absolute error on the test set was 102 s. Out of the four methods, the ms2pip-set shows the lowest error between predicted and experimental RT; compared with MS-GF+ (~186 s), it leads to a decrease of 45% (down to ~102 s, the same as the test error). Both-set and percolator-set lead to a similar decrease in the error to 124 and 125 s, respectively (about 33% less than with MS-GF+).

In Supplementary Materials S8 and S9, the results are presented at the peptide level, where it can be observed that there is a percentage of additional unique peptide identifications of a ratio identical to the additions at the PSM level. Furthermore, the additional peptides map to proteins for which previous evidence was found by MS-GF+ by itself.

To further test and compare the performance of these three feature sets, we show the results of two additional experiments.

3.2 Adult adrenal gland

The proposed framework was evaluated on a larger dataset, downloaded from PRIDE under accession number PXD000561. As mentioned previously, the samples corresponding to an adult adrenal gland contained were selected from this project, and the spectra were searched as described in Supplementary Material S4.

Figure 5 shows how the amount of PSMs obtained by each method increases with the increase of allowed FDR. At 1% FDR, MS-GF+ reports the identification of 23 389 PSMs. This number increases to 24 745 when Percolator's default feature set is used. Using the MS2PIP feature set, the number of PSMs reported at the same level of FDR is 24 013. The largest increase in number of reported PSMs is seen when using the complete feature set; in that case, 25 169 PSMs are reported at 1% FDR, which translates to a 7% increase. This analysis confirmed the previously observation that using the complete feature set is what returns more PSMs at a controlled FDR.

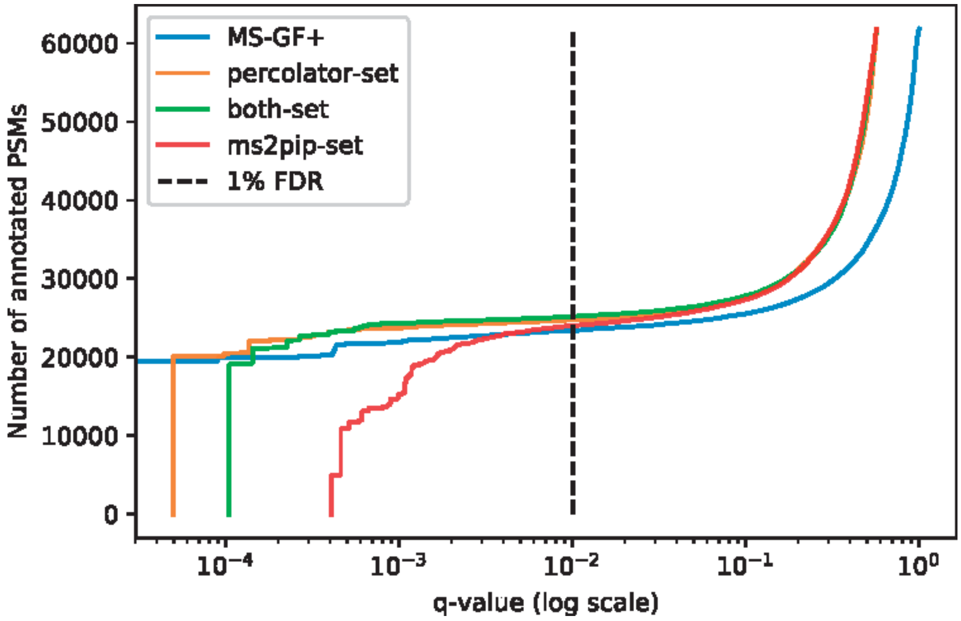

Figure 5: Number of identified PSMs at different levels of FDR for each feature set and for the search engine by itself for the adrenal gland dataset

The overlap between these three post-processors and the original search engine results can be seen in Figure 6. To each feature set, three bars are associated: the orange bars represent the overlap between the rescored set of PSMs and the initial set obtained by MS-GF+; the other two bars represent the PSMs unique to each method (the search engine in blue, i.e. the 'overruled' PSMs, and post-processor in green—i.e. the 'new' PSMs). The overlap between search engine and post-processor is consistently very high. It is noteworthy that ms2pip-set is the feature set which leads to more PSMs previously identified by MS-GF+ to be overruled.

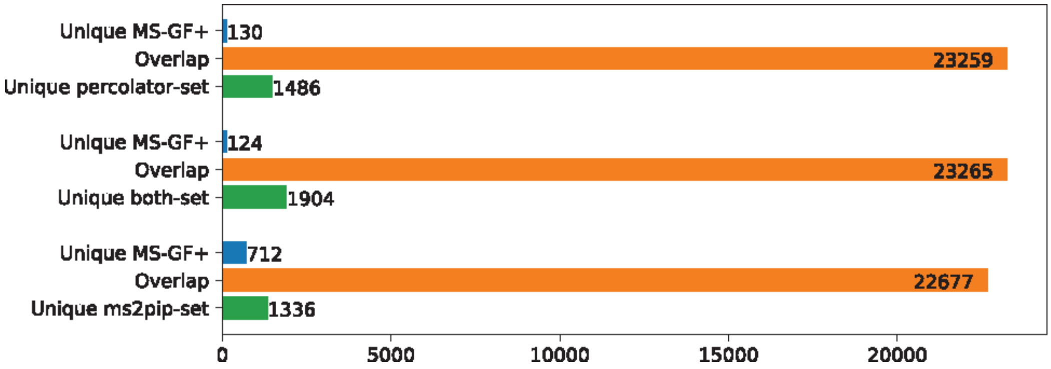

Figure 6: Number and overlap of identified PSMs (1% FDR) on the adrenal gland dataset for each feature set, each compared with MS-GF+ as the baseline. Out of each set of three bars, the first correspond to PSMs obtained uniquely by the search engine, the middle corresponds to the overlap between the post–processor and the search engine, and the bottom bar the PSMs uniquely obtained by the post-processor

An additional analysis can be done concerning the 'new' PSMs obtained with each feature set. In principle, one would expect that if the three feature sets pick up truthful PSMs, then these sets of 'new' PSMs should show significant overlap. This is visualized in the Venn diagram in Figure 7.

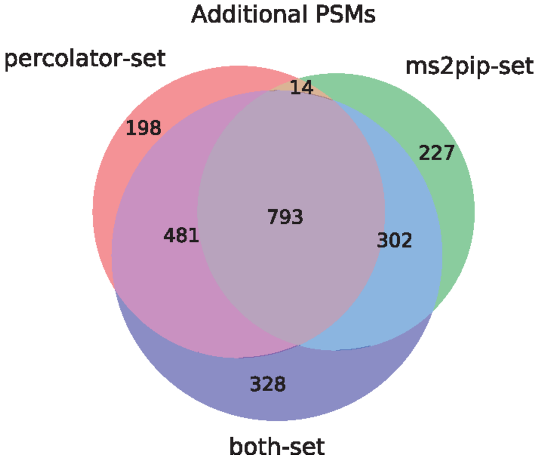

Figure 7: Relationship between the sets of PSMs obtained from the adrenal gland dataset that each feature set adds to the results at 1% FDR

Out of between 1359 and 1883 'new' PSMs, the three features sets agree in 793 of them. Overall, in the case of the default Percolator feature set, about 86% of the added PSMs are supported by at least one other feature set; the same value is 82% for the MS2PIP feature set and 83% when using the combined feature set. This shows a lot of concordance between the three post-processors. As with the previous experiments, the focus of our tests is at the PSM level, but results at the peptide and protein level can be found in Supplementary Material S10.

3.3 HEK sample

For an assessment of performance on a much larger dataset (1 121 149 spectra), the data for accession number PXD001468 were downloaded from the PRIDE database, which was processed as described in Section 2.

As before, Figure 8 shows the increase in PSMs with the increase in allowed FDR. At 1% FDR, 446 336 PSMs are reported by MS-GF+. Using Percolator's default features set increased this number to 486 954, thus adding a little more than 9%. When using only the MS2PIP features, the PSM number increases to 477 324, which is an increase of about 7%. It should be noted that, as was the case for the two datasets discussed above, the highest number of PSMs (495 318, about 11%) is obtained when the default Percolator features are combined with the MS2PIP features. The overlap between the post-processing results for the two feature sets and MS-GF+ alone are shown in Figure 9.

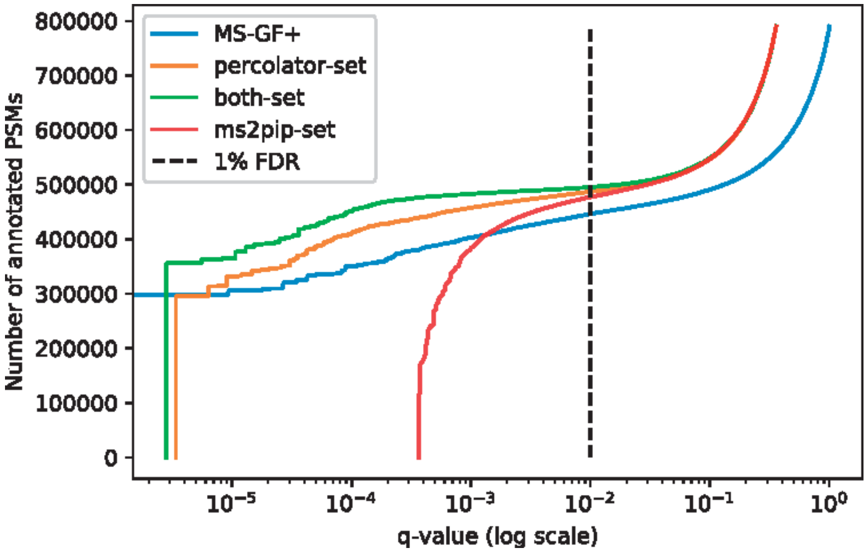

Figure 8: Number of identified PSMs at different levels of FDR for each feature set and for the search engine by itself for the HEK sample dataset

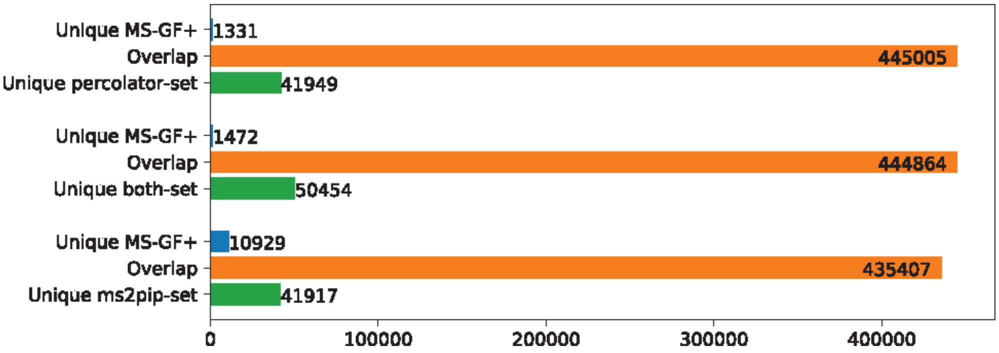

Figure 9: Number and overlap of identified PSMs (1% FDR) on the HEK dataset for each feature set, each compared with MS-GF+ as the baseline. Out of each set of three bars, the first correspond to PSMs obtained uniquely by the search engine, the middle corresponds to the overlap between the post-processor and the search engine, and the bottom bar the PSMs uniquely obtained by the post-processor

As before, the overlap in PSMs added by the three post-processors can be seen to show a lot of agreement between the added PSMs, as can be observed by the plot in Figure 10.

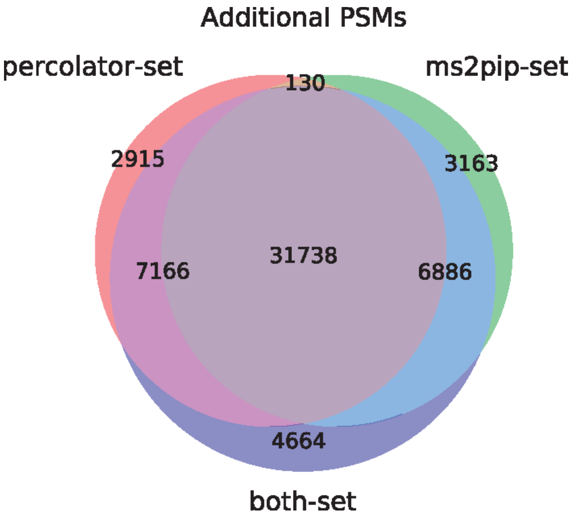

Figure 10: Relationship between the sets of PSMs obtained from the HEK dataset that each feature set adds to the results at 1% FDR

The results of this experiment at the peptide and protein level can be found in Supplementary Material S11.

4 CONCLUSION

In this work, we have showed that, by comparing experimental spectra to theoretical spectra with fragment ion intensities computed by MS2PIP, Percolator can be applied independently from the search engine scoring function. Moreover, we showed that the resulting performance is on par with standard Percolator performance.

Therefore, we have showed that the post-processing step can effectively be decoupled from the search engine used to initially process the data: the only input this post-processing approach requires is a list of target and decoy PSMs, which can be obtained from any search engine or identification strategy. The entire pipeline can be visualized in Supplementary Material S12.

Furthermore, we show consistently that the addition of our MS2PIP feature set to Percolator's default feature set improves the amount of reported PSMs at a controlled level of FDR. As new MS2PIP models are developed, this pipeline can be applied to more specific experiments (such as labeled experiments, or focusing on modifications such as phosphorylation).

As a result, our study emphasizes how computationally predicted spectra can be used to replace 'static' scoring functions with performant and adaptable machine learning algorithms.

FUNDING

This work was supported by SBO grant 'InSPECtor' [120025] of Flanders Innovation and Entrepeneurship (VLAIO), by the European Union's Horizon 2020 Program under Grant Agreement 634402 [PHC32-2014], and by the Research Foundation—Flanders (FWO) [Project No. G.0425.18N]. R.B. received funding from the MASSTRPLAN Marie Sklodowska-Curie EU Framework for Research and Innovation Horizon 2020, under [Grant Agreement No. 675132]. S.D. and L.M. received funding from the European Union's Horizon 2020 Program under Grant Agreement 823839 [H2020-INFRAIA-2018-1].

Conflict of Interest: none declared.

REFERENCES

1. Aebersold R., Mann M. (2003) Mass spectrometry-based proteomics. Nature , 422, 198–207.
2. Anderson D.C. et al. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome Res ., 2, 137–146.
3. Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B , 57, 289–300.
4. CompOmics. (2019a) MS2PIP. https://github.com/compomics/ms2pip_c
5. CompOmics. (2019b) ReScore. https://github.com/compomics/ms2rescore.
6. Chick J. (2015) PRIDE project PXD001468. https://www.ebi.ac.uk/pride/archive/projects/pxd001468
7. Degroeve S., Martens L. (2013) MS2PIP: a tool for MS/MS peak intensity prediction. Bioinformatics , 29, 3199–3203.
8. Eng J.K. et al. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom ., 5, 976–989.
9. Eriksson J. et al. (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. Anal. Chem ., 72, 999–1005.
10. Gabriels R. et al. (2019) Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. Nucleic Acids Res., doi: org/10.1093/nar/gkz299.
11. Käll L. et al. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods , 4, 923–925.
12. Keller A. et al. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem ., 74, 5383–5392.
13. Kim M.-S. et al. (2014) PRIDE Project PXD000561. https://www.ebi.ac.uk/pride/archive/projects/PXD000561.
14. Kim S., Pevzner P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun ., 5.
15. Moruz L. et al. (2010) Training, selection, and robust calibration of retention time models for targeted proteomics. J. Proteome Res ., 9, 5209–5216.
16. Muth T. et al. (2015) Navigating through metaproteomics data: a logbook of database searching. Proteomics , 15, 3439–3453.
17. Nesvizhskii A.I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics , 73, 2092–2123.
18. Perkins D.N. et al. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis , 20, 3551–3567.

19. Sticker A. et al. (2017) Mass spectrometrists should search for all peptides, but assess only the ones they care about. Nat. Methods , 14, 643–644.
20. The M. et al. (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. J. Am. Soc. Mass Spectrom ., 27, 1719–1727.
21. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res ., 45, D158–D169.
22. CrossrefPubMed
23. Vaudel M. (2014) PRIDE project PXD001077. https://www.ebi.ac.uk/pride/archive/projects/pxd001077.
24. Vaudel M. et al. (2012) A complex standard for protein identification, designed by evolution. J. Proteome Res ., 11, 5065–5071.
25. Vaudel M. et al. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat. Biotechnol ., 33, 22–24.
26. Verheggen K. et al. (2017) Anatomy and evolution of database search engines – a central component of mass spectrometry based proteomic workflows. Mass Spectrom. Rev., 2017, 1–15.
27. Vizcaíno J.A. et al. (2016) 2016 update of the PRIDE database and its related tools. Nucleic Acids Res ., 44, D447–D456.
28. Yilmaz S. et al. (2017) Methods to calculate spectrum similarity. In: Keerthikumar S., Mathivanan S. (eds), Proteome Bioinformatics , Vol. 1549. Springer New York, New York, NY, pp. 75–100.