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

Ana S. C. Silva,[*,†,‡,¶] Lennart Martens,[*,†,‡,¶] and Sven Degroeve[*,†,‡,¶]

†*VIB-UGent Center for Medical Biotechnology, Ghent, 9000, Belgium*

‡*Department of Biochemistry, Faculty of Medicine, Ghent University, Ghent, 9000, Belgium*

¶*Bioinformatics Institute Ghent, Ghent University, Ghent, 9000, Belgium*

E-mail: anascsilva@vib-ugent.be; lennart.martens@vib-ugent.be; sven.degroeve@vib-ugent.be

Abstract

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. By leveraging another machine learning-based tool, MS2PIP, we aim to overcome this obstacle. MS2PIP predicts fragment ion peak intensities. We show how comparing these intensities to annotated experimental spectra by calculating direct similarity metrics rather than the more common peak counting or explained intensities summing provides enough information for a tool such

as Percolator to accurately separate two classes of PSMs, recovering more information out of the data while maintaining control of statistics such as the false discovery rate.

Introduction

Proteomics is a field that relies heavily on mass spectrometry for the identification of proteins in a sample.[1] 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.[2]

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.[2] These theoretical spectra are typically very simple: for example, SEQUEST[3] creates these by assigning an arbitrary magnitude of 50 for b- and y-ion fragment peaks, of 25 for ions with m/z equal to +-1u 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. Today, these scoring functions are typically extended with a statistical model that attempts to model the probability of obtaining a match at this score by chance.[4] The output of a search engine thus 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.

While 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 very 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[5].[6]

In proteomics, FDR control is not based on statistical models, but is 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, while the decoy database contains nonsensical sequences that are designed to accurately represent false positive identification results.[7]

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. Table 1 in the SI lists the features used in this research. Post-processing tools have been available from as early as 2002,[8] 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, while the negative class is represented by decoy PSMs. One such attempt was Anderson et al.'s,[9] who trained a support vector machine (SVM) on PSMs obtained from SEQUEST.[3] 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, while the the negative class corresponding to PSMs associated with incorrect proteins.

Percolator[10]perfects this approach. While similar in principle, it introduces some crucial

differences that have made it into the most popular post-processing tool in the proteomics community.[11] The first of these differences was the re-framing of the problem as a semi-supervised problem: 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 classes. The second crucial difference is that this new scoring function is not meant to generalize to other data sets, but only to rescore the PSMs in the current data set - including the PSMs in the "unknown" class. In this way, members of this class can be re-assigned to the positive class, and more PSMs can be recovered at the same FDR cut-off.

Here, we overcome the search engine dependency of post-processing algorithms by providing ReScore, a tool that leverages fragmentation spectrum intensity predictions by MS2PIP[12] to provide additional discriminative and complementary features to post-process PSMs using machine learning. By comparing the intensity information in these predicted spectra with those recorded in the experimental spectra, ReScore can be used to sensitively identify truthful PSMs while maintaining controlled FDR. This allows ReScore to go beyond trivial 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 ReScore effectively independent of the original search engine score and related features, thus making ReScore compatible with any desired search engine, or with the combined output of multiple search engines, as for instance obtained by PeptideShaker.[13]

Methods

MS2PIP

MS2PIP is used to predict fragment ion intensities given a (modified) peptide sequence and charge. MS2PIP contains different models for different fragmentation types; definite models for singly charged fragment ions exist for two fragmentation types: HCD and CID. Models for EThcD fragmentation and for doubly charged fragment ions in HCD are also available in beta. Since its initial publication, MS2PIP has gone through several updates, and is currently available as a web-service[14 15] and as a locally installable package.[16]

The input for MS2PIP is a PEPREC (PEPtide RECord[16]) 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[17]), it outputs both the experimental intensity (obtained from the mgf file) and the predicted intensity for each fragment ion.

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 contain log2 transformed total-ion-current (TIC) normalized intensities for the theoretical fragment ions[12].[18] Given the nature of the log_2 transformation, low intensity peaks are thus weighed relatively more heavily by spectrum comparison metrics. Spectral correlation features are therefore calculated twice: once on the log2 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 fragment

ion 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 SI, T1. Note that this set of features will include significant redundancy, but this can be overcome in an unbiased, data-driven way by the application of feature selection routines described below.

To reduce the wide set of initial features listed in SI T1, data-driven methods are used to *a posteriori* select a smaller subset of features that allows for the classification task to be carried out with good performance.[19] This selection of the most relevant features is done by analysing the weight parameters that the machine learning model (in the case at hand an SVM) learns, which reflect which features are more and which features are less relevant for the current task. A new model is re-trained on the subset of features with the highest absolute weights, and its performance is compared to its counterpart which had been trained on the complete set of features. This results in the removal of noisy or redundant features, which reduces the complexity of the problem and frequently leads to an improvement of the model.

Data sets and processing

A test data set for the feature selection was compiled based on the established *Pyrococcus furiosus* (Pfu) standard for protein identification was used,[20] as obtained from the PRIDE archive[21].[22] The spectra were searched with MS-GF+[23] against a FASTA[24] formatted sequence database that includes all Pfu protein sequences (both reviewed and unreviewed), obtained via UniProt.[25] More details on the search settings can be found in SI 1.

The search results were converted into a Percolator input file (pin),[10] 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+). After running percolator with this file, the initial output of

9795 PSMs identified at 1% FDR by MS-GF+ alone, was improved to 10507 PSMs identified at 1% FDR by MS-GF+/Percolator. Additionally to a regular search with this dataset, we perform the entrapment experiment as proposed in.[20] This is an experiment consists of adding the human proteome to the database against which the Pfu spectra are searched. These sequences are called entrapment sequences. We can use them to validate the FDR calculation using decoy sequences, by confirming how many of these entrapment sequences pass the score threshold and get into our results set. Details can be found in SI 2.

An input file for MS2PIP (that is, a PEPREC file) was created. This file includes all the unfiltered PSMs obtained by a search engine (in this case MS-GF+), and it 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 table in SI T1. 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 ran 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.

This procedure was repeated for two additional larger and more noisy datasets, both obtained from PRIDE. The first contains two types of human tissue (brain and kidney),[26] and the second is a selection from the draft of the human proteome[27] – specifically, all datasets from an adult adrenal gland.[28] Details on the search settings used can be found in SI 3 and SI 4, respectively.

The code and data necessary to reproduce these results are available on github,[29] along with installation and usage instructions.

Results and discussion

Pyroccocus furiosus (Pfu) dataset

The discriminative power of the spectral intensity prediction features computed by ReScore 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, only MS2PIP-derived intensity correlation features, and the Percolator and MS2PIP-derived features combined. For each comparison we looked at the overlap in PSM identifications.

Figure 1 shows the outcome of these comparisons, with each set of three bars representing one comparison. All results are reported at a fixed FDR level of 1%.

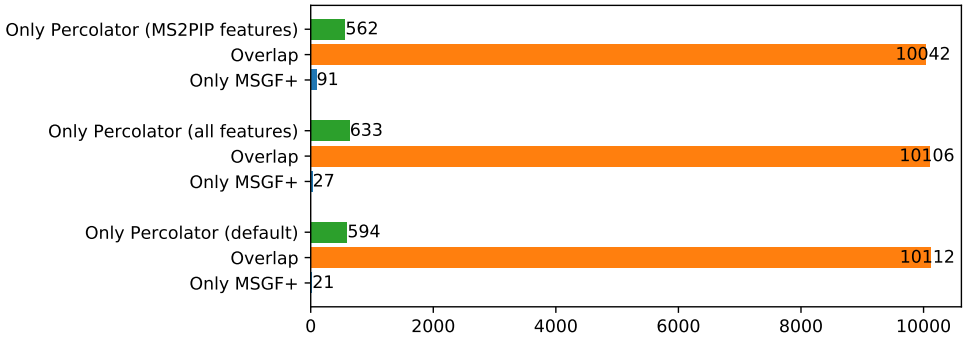


Figure 1: Number and overlap of identified PSMs (1% FDR) for each feature set, each compared with MS-GF+ as the baseline. Orange bars represent the overlap between post-processed PSM results and the MS-GF+ PSM results, green bars represent the amount of PSMs uniquely obtained by the post-processing method, while blue bars represent PSMs uniquely obtained by the search engine.

As expected the PSM sets identified by the different post-processing methods maintain a consistently large overlap with the results obtained by MS-GF+ alone, and moreover all post-processing methods contribute a very similar number of additional identifications. This initial result shows that the information computed from the MS2PIP predictions alone, when processed by Percolator, is at least as discriminating as the search engine's scoring function, making it possible to post-process PSMs completely independently of the search

engine algorithm and score. In fact, the only relevant input to ReScore in this scenario is a spectrum and its proposed peptide match. It is also noteworthy that this scenario "overrules" more MS-GF+ matches than the other two post-processing feature sets.

Figure 2 shows the overlap between the additional PSMs that each of the three different post-processing feature sets reported when compared against only MS-GF+ (represented by the green bars in Figure 1). While more than half of these additional PSMs are agreed upon by all post-processing feature sets, and at least a third of these additional PSMs is shared by at least two feature sets, a fair amount of disagreement remains.

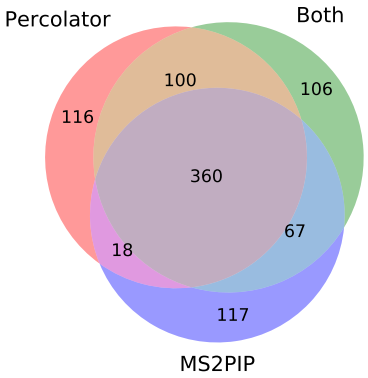


Figure 2: Overlap between additional PSMs picked up by the different post-processing feature sets at 1% FDR when compared with MS-GF+ alone.

The scoring functions that Percolator learns are essentially linear combinations of the features it is provided with. By extracting the weights that Percolator gives to each feature it is therefore possible to study the importance of this feature to the task of separating real from false PSMs. The representation of these weights across Percolator's internal cross-validation folds (Figure 3) shows that several of Percolator's default features do not contribute at all to the learned scoring function (specifically, the amino-acid frequencies seen in the top left side of the figure, which have a weight of 0, meaning they are essentially discarded by the learned scoring function).

Indeed, final search engine score (feature "lnEValue") and peptide length are the two main

features by weight in the Percolator default feature set, whereas we see many more highly weighed features in the MS2PIP feature set.

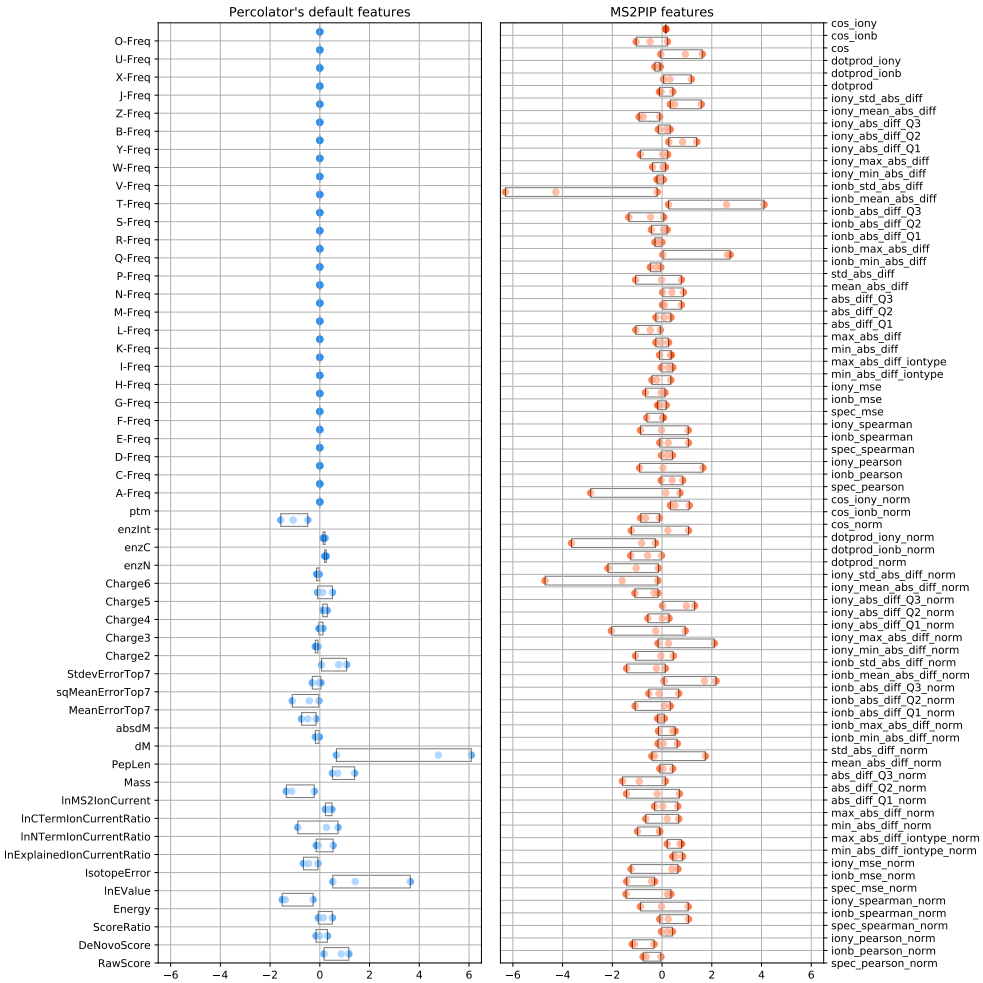


Figure 3: Feature weights Percolator uses to build the new scoring function.

In order to establish independence of the search engine scoring function, we then eliminated all Percolator default features. Moreover, because the relative feature weights for the MS2PIP only feature set were very similar to these for the combined feature set (see SI 6), we then proceeded to also remove the MS2PIP only features with low weights, obtaining a final set of 25 features. This set of features is shown in SI 5. Percolator was ran with this reduced file as input, and, as Figure 4 shows, was still able to accurately separate additional PSMs at 1% FDR.

While the overlap in obtained PSMs provides a reasonable indicator of their quality, a

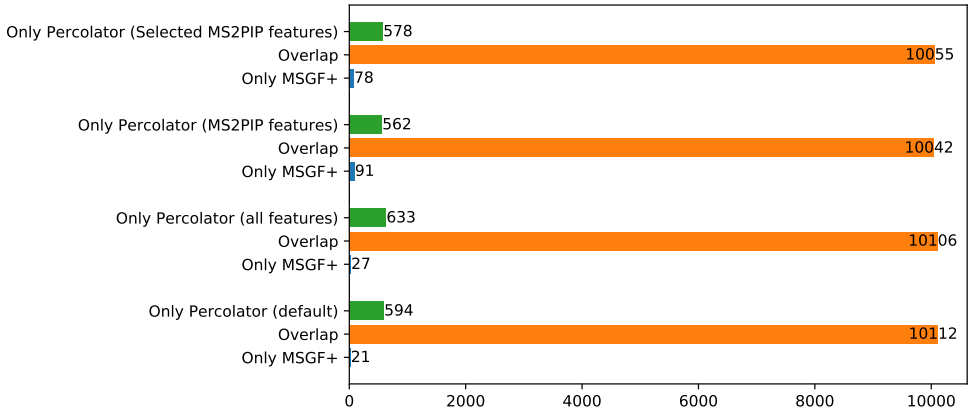
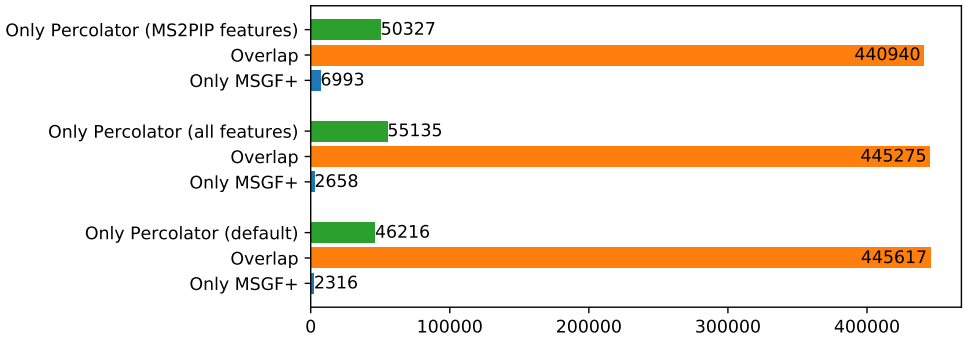


Figure 4: Number and overlap of identified PSMs (1% FDR) for each feature set, including the reduced MS2PIP feature set, each compared with MS-GF+ as the baseline. Orange bars represent the overlap between post-processed PSM results and the MS-GF+ PSM results, green bars represent the amount of PSMs uniquely obtained by the post-processing method, while blue bars represent PSMs uniquely obtained by the search engine

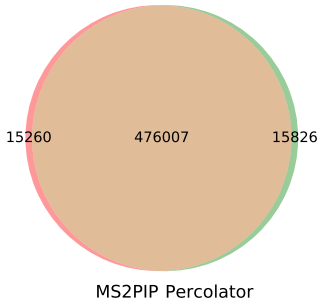
more in-depth analysis on identification reliability can be performed by adding entrapment sequences to the Pyrococcus furiosus search. These entrapment sequences can then be used as a direct way to assess whether the rescoring models have overfit to the data. The percentage of entrapment sequences in the results reported by MS-GF+ alone is 0.23%, which is well below the reported 1% FDR. The percentage of entrapment sequences for the Percolator default feature set is comparable at 0.22%. When using only the MS2PIP feature set, however, the percentage of entrapment sequences drops by a quarter to 0.16%. From this result, we can infer that a substantial part of the PSMs that are discarded when using the MS2PIP feature set are likely low quality or incorrect PSMs.

Homo sapiens data-set

For an assessment of ReScore performance on a much larger data set, the data for accession number PXD001468[26] was downloaded from the PRIDE database, which was processed as described in the methods section.



(a)



(b)

Figure 5: Amount of PSMs obtained with MS-GF+ alone, with default Percolator feature set, with only the MS2PIP feature set and with the joint feature set (a), and overlap in PSMs identified by the two post processing methods (b)

Adult Adrenal Gland

The proposed framework was evaluated in one more dataset, downloaded from PRIDE under accession number PXD000561. As mentioned previously, one of the several tissues included in this experiment was selected and the spectra were searched as described in SI 4.

At 1% FDR, MS-GF+ reports the identification of 23521 PSMs. This number increases to 24995 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 24851 – in both cases, an increase of about 6%. As seen previously, the largest increase in number of reported PSMs is seen when using the complete feature set; in that case, 25590 PSMs are reported at 1% FDR which

translates an 8% increase. This analysis confirmed the previously observation that using the complete feature set is what returns more PSMs at a controlled FDR. The overlap between these three cases and the original search engine results can be seen in Fig. 6.

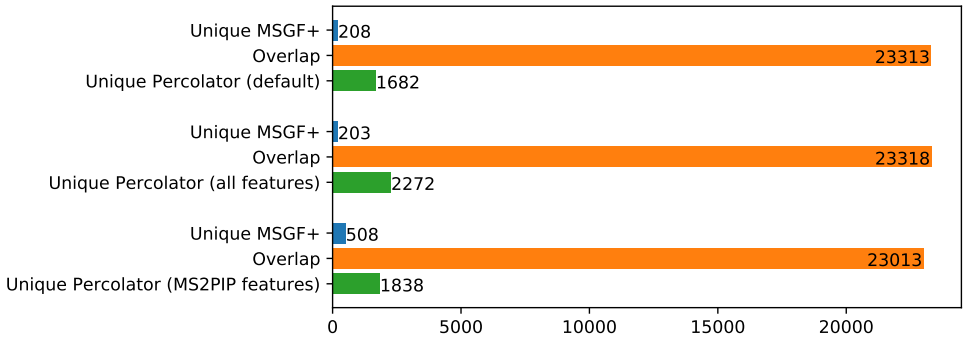


Figure 6: Amount of PSMs obtained with MS-GF+ alone, with default Percolator feature set, with only the MS2PIP feature set and with the joint feature set.

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

Acknowledgement

This work was supported by SBO grant "InSPECtor" (120025) of Flanders Innovation and Entrepeneurship (VLAIO), by the European Union's Horizon 2020 Programme under Grant Agreement 634402 (PHC32-2014), and by the Research Foundation – Flanders (FWO) (Project No. G.0425.18N).

Supporting Information Available

The following is available in the supporting information:

- Table 1: Spectral comparison features
- 1: details on the Pfu normal search
- 2: details on the Pfu entrapment search
- 3: details on the *Homo sapiens* brain and kidney search
- 4: details on the *Homo sapiens* adrenla gland search
- 5: selected MS2PIP features
- 6: relative feature weights

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. **2003**, *422*, 10.

(2) Verheggen, K.; Ræder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews* **2017**, *0*.

(3) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994**, *5*, 976–989.

(4) Yilmaz, S.; Vandermarliere, E.; Martens, L. In *Proteome Bioinformatics*; Keerthikumar, S., Mathivanan, S., Eds.; Springer New York: New York, NY, 2017; Vol. 1549; pp 75–100.

(5) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* **1995**, *57*, 289–300.

(6) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* **2010**, *73*, 2092–2123.

(7) Sticker, A.; Martens, L.; Clement, L. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nature Methods* **2017**, *14*, 643–644.

(8) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry* **2002**, *74*, 5383–5392.

(9) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. A New Algorithm for the Evaluation of Shotgun Peptide Sequencing in Proteomics: Support Vector Machine Classification of Peptide MS/MS Spectra and SEQUEST Scores. *Journal of Proteome Research* **2003**, *2*, 137–146.

(10) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **2007**, *4*, 923–925.

(11) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of The American Society for Mass Spectrometry* **2016**, *27*, 1719–1727.

(12) Degroeve, S.; Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **2013**, *29*, 3199–3203.

(13) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology* **2015**, *33*, 22–24.

(14) Degroeve, S.; Maddelein, D.; Martens, L. MS ^{2} PIP prediction server: compute and visualize MS ^{2} peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research* **2015**, *43*, W326–W330.

(15) Gabriels, R.; Turan, D.; Degroeve, S. MS2PIP Server. `https://iomics.ugent.be/ms2pip/`.

(16) Degroeve, S. MS2PIP. `https://github.com/compomics/ms2pip_c`.

(17) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(18) Degroeve, S.; Colaert, N.; Vandekerckhove, J.; Gevaert, K.; Martens, L. A reproducibility-based evaluation procedure for quantifying the differences between MS/MS peak intensity normalization methods. *PROTEOMICS* **2011**, *11*, 1172–1180.

(19) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389–422.

(20) Vaudel, M.; Burkhart, J. M.; Breiter, D.; Zahedi, R. P.; Sickmann, A.; Martens, L.

A Complex Standard for Protein Identification, Designed by Evolution. *Journal of Proteome Research* **2012**, *11*, 5065–5071.

(21) Vizcaíno, J. A.; Csordas, A.; del Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research* **2016**, *44*, D447–D456.

(22) Vaudel, M. Project PXD001077. 2014; `https://www.ebi.ac.uk/pride/archive/projects/PXD001077`.

(23) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **2014**, *5*.

(24) DJ Lipman,; WR Pearson, Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435–1441.

(25) The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **2017**, *45*, D158–D169.

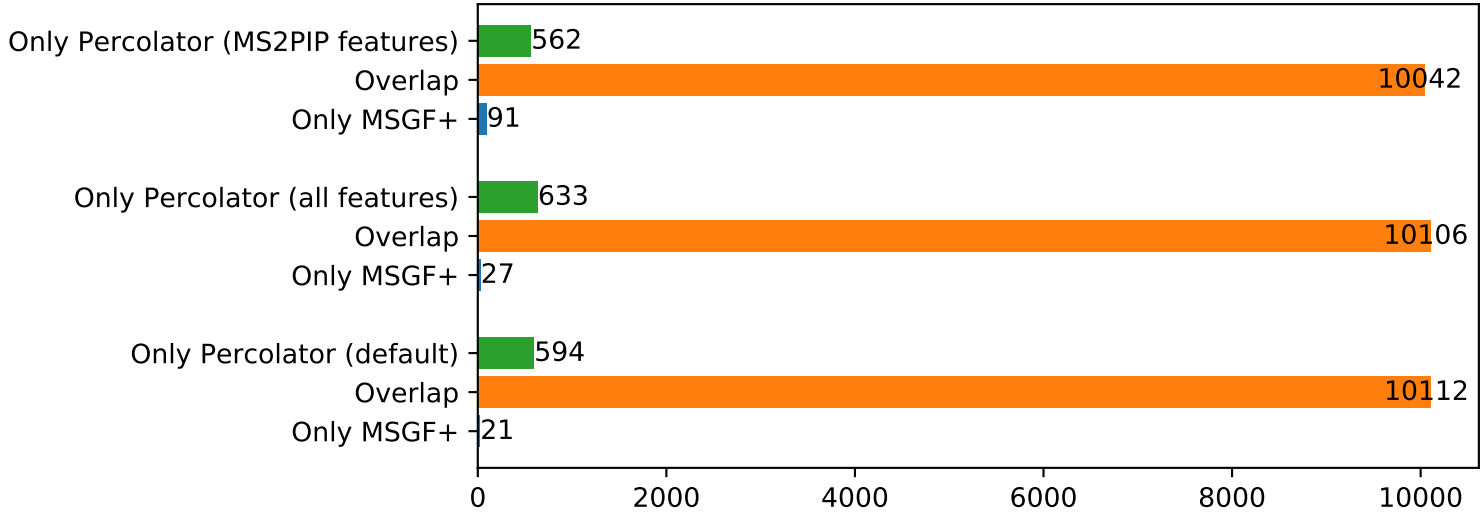
(26) Chick, J. Project PXD001468. 2015; `https://www.ebi.ac.uk/pride/archive/projects/PXD001468`.

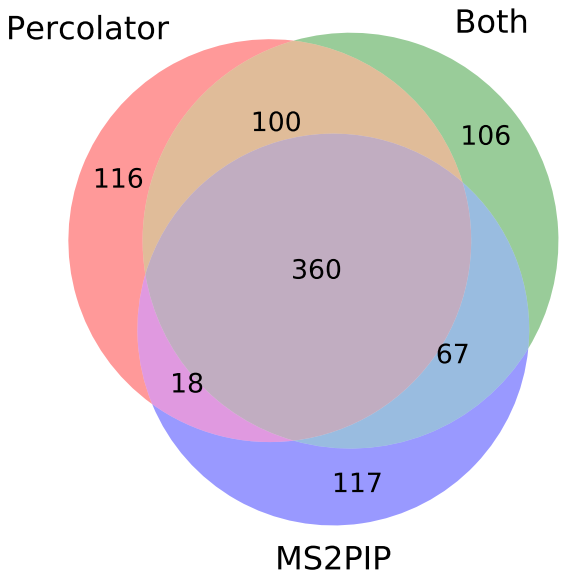
(27) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T.-C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.;

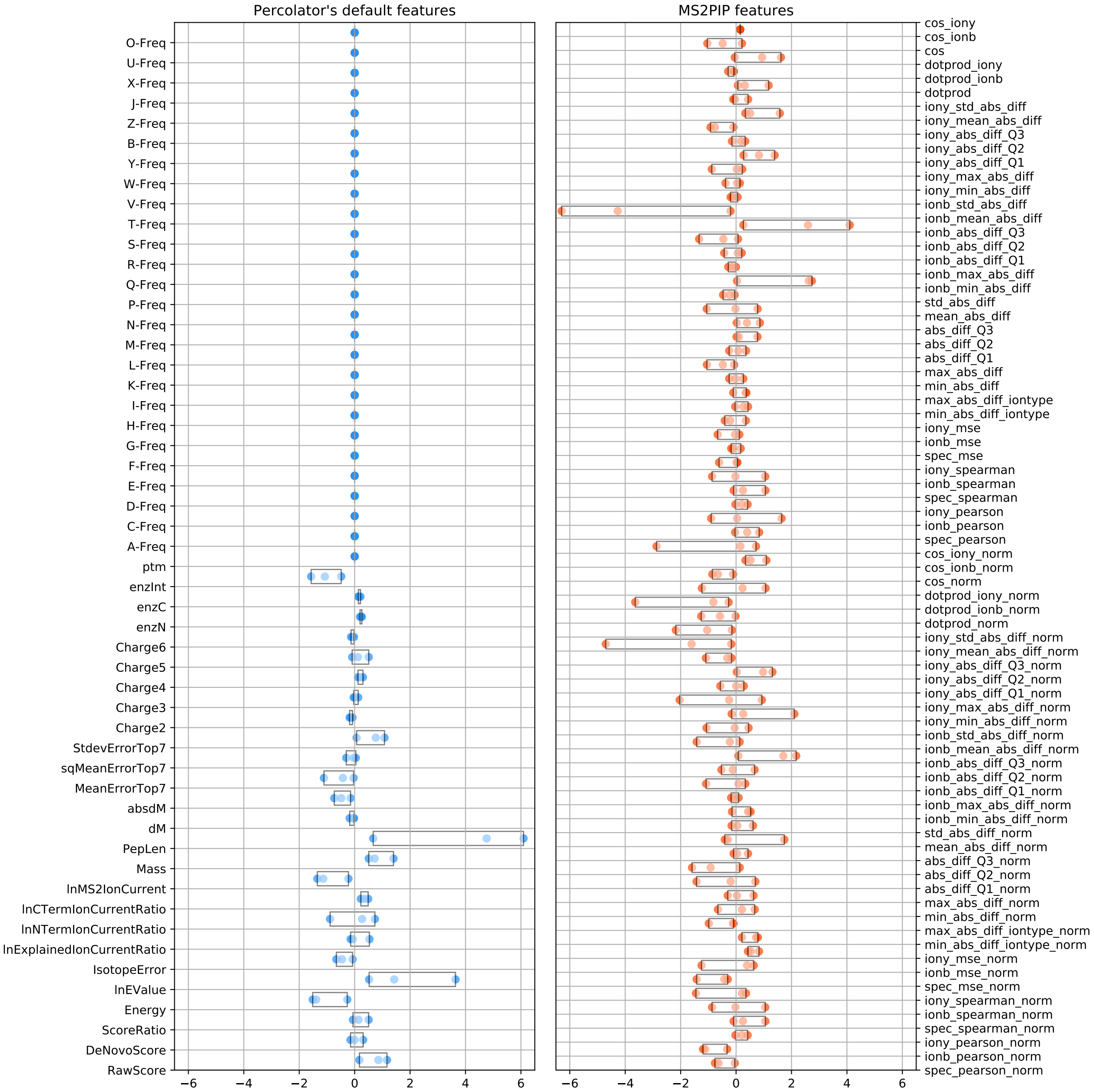
Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S. K.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509*, 575–581.

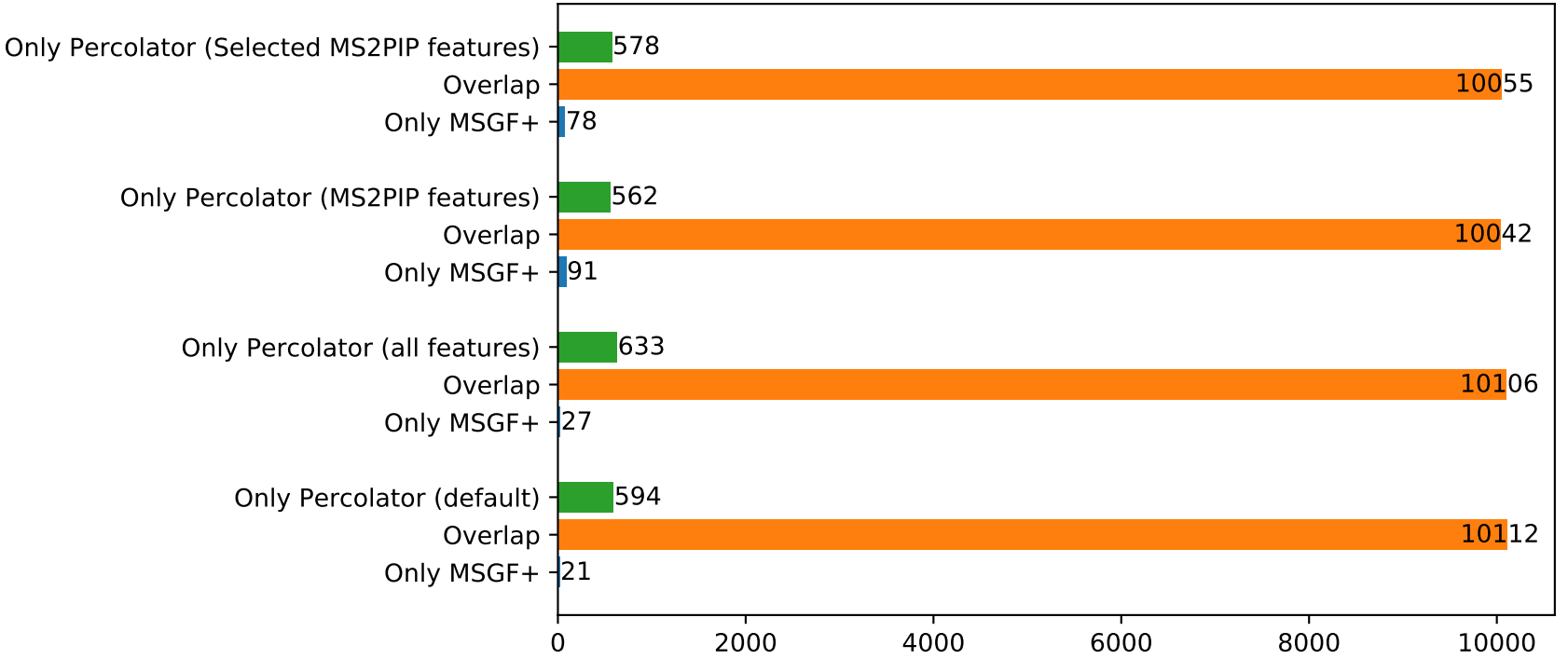
(28) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T.-C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S. K.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. Project PXD000561. 2014; `https://www.ebi.ac.uk/pride/archive/projects/PXD000561`.

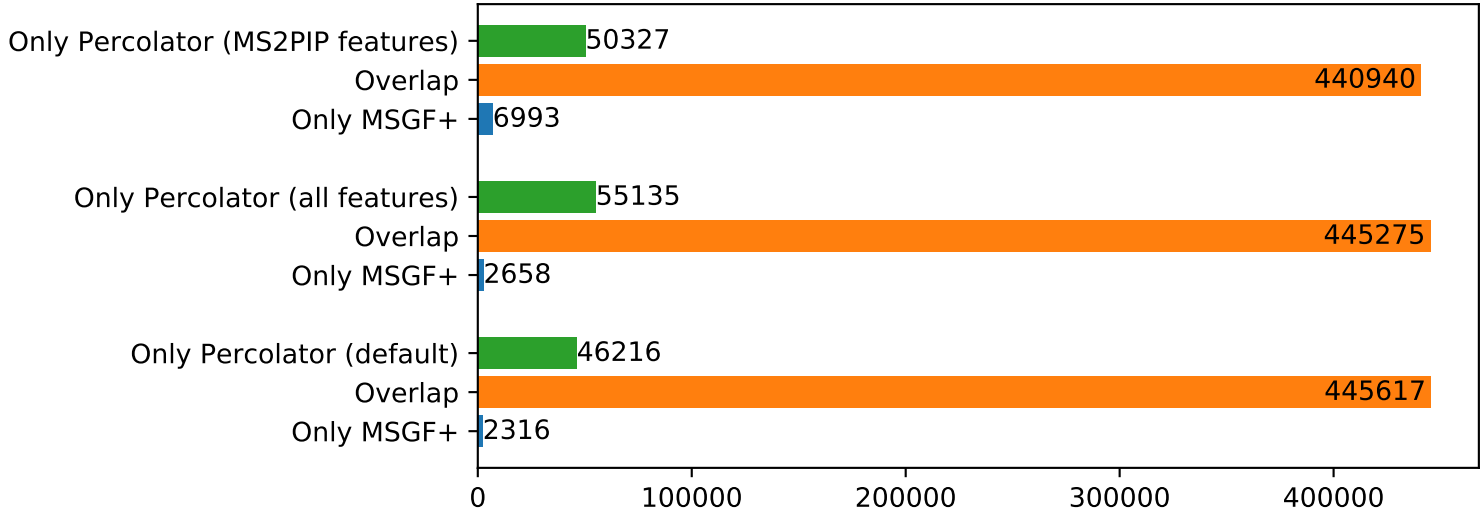
(29) C. Silva, A. S. rescore. `https://github.com/anasilviacs/rescore`.

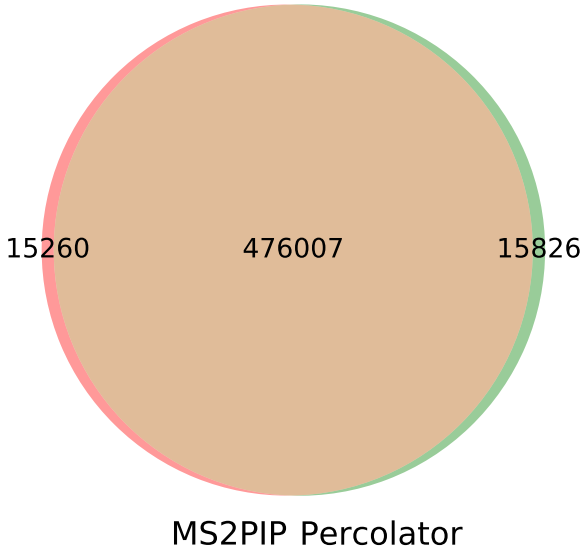

Only Percolator (MS2PIP features)
Overlap
Only MSGF+
Only Percolator (all features)
Overlap
Only MSGF+
Only Percolator (default)
Overlap
Only MSGF+
562
10042
91
633
10106
27
594
10112
21
0
2000
4000
6000
8000
10000

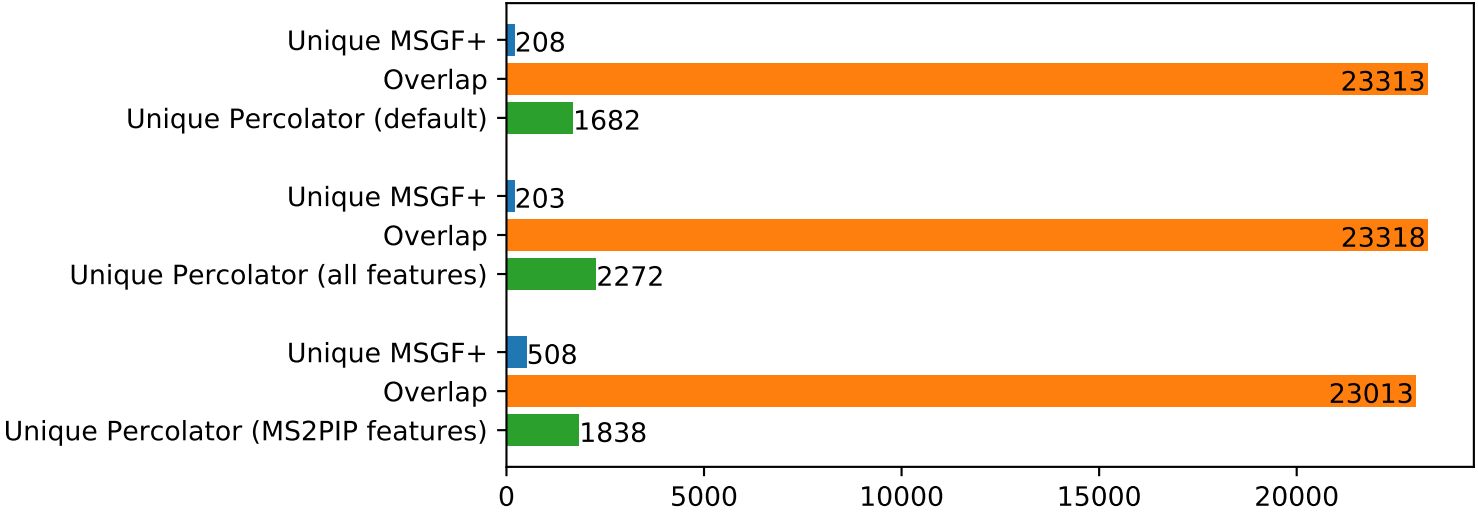

Percolator
Both
100
106
116
360
67
18
117
MS2PIP


Percolator's default features
MS2PIP features
O-Freq
U-Freq
X-Freq
J-Freq
Z-Freq
B-Freq
Y-Freq
W-Freq
V-Freq
T-Freq
S-Freq
R-Freq
Q-Freq
P-Freq
N-Freq
M-Freq
L-Freq
K-Freq
I-Freq
H-Freq
G-Freq
F-Freq
E-Freq
D-Freq
C-Freq
A-Freq
ptm
enzInt
enzC
enzN
Charge6
Charge5
Charge4
Charge3
Charge2
StdevErrorTop7
sqMeanErrorTop7
MeanErrorTop7
absdM
dM
PepLen
Mass
lnMS2IonCurrent
lnCTermIonCurrentRatio
lnNTermIonCurrentRatio
lnExplainedIonCurrentRatio
IsotopeError
lnEValue
Energy
ScoreRatio
DeNovoScore
RawScore
−6
−4
−2
0
2
4
6
cos_iony
cos_ionb
cos
dotprod_iony
dotprod_ionb
dotprod
iony_std_abs_diff
iony_mean_abs_diff
iony_abs_diff_Q3
iony_abs_diff_Q2
iony_abs_diff_Q1
iony_max_abs_diff
iony_min_abs_diff
ionb_std_abs_diff
ionb_mean_abs_diff
ionb_abs_diff_Q3
ionb_abs_diff_Q2
ionb_abs_diff_Q1
ionb_max_abs_diff
ionb_min_abs_diff
std_abs_diff
mean_abs_diff
abs_diff_Q3
abs_diff_Q2
abs_diff_Q1
max_abs_diff
min_abs_diff
max_abs_diff_iontype
min_abs_diff_iontype
iony_mse
ionb_mse
spec_mse
iony_spearman
ionb_spearman
spec_spearman
iony_pearson
ionb_pearson
spec_pearson
cos_iony_norm
cos_ionb_norm
cos_norm
dotprod_iony_norm
dotprod_ionb_norm
dotprod_norm
iony_std_abs_diff_norm
iony_mean_abs_diff_norm
iony_abs_diff_Q3_norm
iony_abs_diff_Q2_norm
iony_abs_diff_Q1_norm
iony_max_abs_diff_norm
iony_min_abs_diff_norm
ionb_std_abs_diff_norm
ionb_mean_abs_diff_norm
ionb_abs_diff_Q3_norm
ionb_abs_diff_Q2_norm
ionb_abs_diff_Q1_norm
ionb_max_abs_diff_norm
ionb_min_abs_diff_norm
std_abs_diff_norm
mean_abs_diff_norm
abs_diff_Q3_norm
abs_diff_Q2_norm
abs_diff_Q1_norm
max_abs_diff_norm
min_abs_diff_norm
max_abs_diff_iontype_norm
min_abs_diff_iontype_norm
iony_mse_norm
ionb_mse_norm
spec_mse_norm
iony_spearman_norm
ionb_spearman_norm
spec_spearman_norm
iony_pearson_norm
ionb_pearson_norm
spec_pearson_norm
−6
−4
−2
0
2
4
6


Only Percolator (Selected MS2PIP features)
Overlap
Only MSGF+
Only Percolator (MS2PIP features)
Overlap
Only MSGF+
Only Percolator (all features)
Overlap
Only MSGF+
Only Percolator (default)
Overlap
Only MSGF+
578
10055
78
562
10042
91
633
10106
27
594
10112
21
0
2000
4000
6000
8000
10000


Only Percolator (MS2PIP features)
Overlap
Only MSGF+
Only Percolator (all features)
Overlap
Only MSGF+
Only Percolator (default)
Overlap
Only MSGF+
50327
440940
6993
55135
445275
2658
46216
445617
2316
0
100000
200000
300000
400000


15260
476007
15826
MS2PIP
Percolator


Unique MSGF+
Overlap
Unique Percolator (default)
Unique MSGF+
Overlap
Unique Percolator (all features)
Unique MSGF+
Overlap
Unique Percolator (MS2PIP features)
208
23313
1682
203
23318
2272
508
23013
1838
0
5000
10000
15000
20000