Sensitive and specific spectral library searching with COSS and Percolator

Genet Abay Shiferaw [1,2], *Ralf Gabriels* [1,2], *Robbin Bouwmeester*[1,2], *Tim Van Den Bossche*[1,2], *Elien Vandermarliere* [1,2], *Lennart Martens* [1,2,*], *Pieter-Jan Volders* [1,2,3]

[1] VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium

[2] Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium

[3] Cancer Research Institute Ghent, Ghent University, 9000 Ghent, Belgium

Corresponding Author

*Prof. Dr. Lennart Martens, Technologiepark-Zwijnaarde 75, 9052 Ghent, Belgium.

E-mail: lennart.martens@vib.ugent.be , Tel: +32 92249840

ABSTRACT: Maintaining high sensitivity while limiting false positives is a key challenge in peptide identification from mass spectrometry data. Here, we therefore investigate the effects of integrating the machine learning-based post-processor Percolator into our spectral library searching tool COSS (CompOmics Spectral library Searching tool). To evaluate the effects of this post-processing, we have used forty data sets from two different projects and have searched these against the NIST and MassIVE spectral libraries. The searching is carried out using two spectral library search tools, COSS and MSPepSearch with and without Percolator post-processing, and using sequence database search engine MS-GF+ as a baseline comparator. The addition of the Percolator rescoring step to COSS is effective and results in a substantial improvement in sensitivity and specificity of the identifications. COSS is freely available as open source under the

permissive Apache2 license, and binaries and source code are found at https://github.com/compomics/COSS

Keywords: spectral library, peptide identification, rescoring, COSS, Percolator, mass spectrometry

Introduction

MS-based peptide identification typically relies on matching measured spectra against theoretical spectra in a database searching approach[1]. However, identification can also be obtained by matching measured spectra against a spectral library consisting of previously measured and identified spectra[2]. Several spectral library searching tools have been developed for this purpose, with notable examples including SpectraST[3], the National Institute of Standards and Technology (NIST) MS Search[4] and MSPepSearch (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:mspepsearch), ANN-SoLo[5], X!Hunter[6], and COSS[7]. The direct comparison of a newly measured spectrum against the spectra in such a spectral library can both increase sensitivity[8] and reduce computational complexity compared to database searching[9].

However, similar to database searching, spectral library searching is not perfect, and several causes can lead to incorrect peptide identification[10]: quality of the spectral data, unexpected post translational modifications, charge state issues, or a poor scoring function. To control this erroneous identification, the validation of search results is a crucial step in the identification process. Typically, this is handled through a target-decoy approach in both database and spectral library searches[1]. This approach allows estimation of the False Discovery Rate (FDR)[11,12], which is the expected proportion of incorrect peptide to spectrum matches (PSMs) among the selected set of accepted identifications.

Besides the validation of PSMs using target-decoy based FDR control, it has also become common to employ post-processing methods to database search results to increase sensitivity. The most popular of these is Percolator[13], which significantly improves the sensitivity of multiple database search engines, including SEQUEST, Mascot[14], and MS-GF+[15]. Percolator itself is a semi-supervised machine learning algorithm based on a linear support vector machine, which is designed to discriminate between correct and incorrect peptide matches by rescoring peptide identifications. For this, Percolator considers a set of features that describe each PSM and uses these features as well as the annotation of target and decoy PSMs in an iterative process to re-rank PSMs using a new score and associated q-value. Yet, although Percolator is commonly used in database[16,17] searches, its use has thus far not been reported for spectral library searching.

In this manuscript, we have examined the utility of Percolator in spectral library searching by extending our COSS spectral library search tool. Importantly, we have found that Percolator improves the output of COSS in two ways. First, the total number of identification increases due to rescoring at 1% FDR. Second, peptides that are misidentified before rescoring are deleted while some peptides that wrongly unidentified are added to the final rescoring result.

We therefore integrated Percolator into the latest version of our freely available, open-source COSS tool, which is moreover capable of handling multiple file formats, and can also be used to analyze large data sets.

MATERIAL AND METHODS

Experimental data sets and spectral library

We obtained raw data files from the deep proteome and transcriptome abundance atlas[18] data set (36 runs corresponding to one fractionated brain sample; ProteomeXchange ID PXD010154) and four runs from "Assessing the relationship between mass window width and retention time scheduling on protein coverage for data-independent acquisition"[19] data set (ProteomeXchange ID PXD013477, data dependent acquisition (DDA) runs of the HeLa sample) as benchmarking data sets (Supplementary Table S-1). All these raw files were converted to Mascot Generic Format (mgf) format using the msconvert tool (ProteoWizard[20] version 3.0.19014), with the peak picking algorithm activated (Figure S-1). The well-known NIST spectral library (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:cdownload, obtained on 15/08/2021) and MassIVE[21] (obtained on 18/09/2018) were used to perform our searches against. More information of these two spectral libraries can be found in Supplementary Table S-2.

Spectral library searches

All 40 data sets are searched with COSS (COSS-2.0) against the NIST and MassIVE spectral library, which were appended with the corresponding decoy spectra generated with the COSS decoy spectra generator with the reverse sequence method[7]. The following search settings were used: precursor mass tolerance set to 10 ppm, fragment mass tolerance set to 0.05 Dalton (Da), MSROBIN scoring function[7], and a fragment mass window to select peaks set to 10 Da.

To run MSPepSearch (version 0.96), we first generated and concatenated an MSPepSearch decoy spectral library using the COSS decoy generator (reverse sequence method). Next, we converted the msp file of this library to MSPepSearch's binary file format using Lib2NIST (version 1.0.6.5) and performed the searches with a precursor mass tolerance of 10 ppm and fragment mass tolerance 0.05 Da. We were unable to successfully perform searches with MSPepSearch against the MassIVE spectral library due to its size and the resulting memory requirements.

Sequence database search

In parallel with the spectral library searches, we performed a sequence database search with the well-established MS-GF+[22] (version 2021.01.08) tool through SearchGUI[23] (version 4.0.22) with additional features reporting activated for use by Percolator. The search database was constructed from the human reference proteome (UP000005640) as obtained from UniProtKB[24] (consulted on 9/10/2018). Carbamidomethylation of cysteine was set as a fixed, and oxidation of methionine as a variable modification. Trypsin was set as protease and a maximum of two missed cleavages was allowed. Precursor mass tolerance was set to 10 ppm, and fragment mass tolerance to 0.05 Da. Precursor charges from 2 to 4 were considered. The SearchGUI configuration file used to run MS-GF+ is provided at https://github.com/compomics/COSS-Percolator-manuscript .

False discovery rate estimation

Validation of the obtained results is a key step in peptide identification to control erroneous results, and typically takes the form of false discovery rate (FDR) control [11]. Different types of decoy spectral library generators are already implemented and added to COSS[7]. For this experiment, we have used decoy type which uses reverse and random sequence decoy generation technique as described in Zhang et al. [25]. Briefly, the sequence of each spectrum is reversed, leaving the last amino acid in place. Based on this sequence, the masses of the a, b and y ions are calculated and the corresponding annotated peaks in the spectrum are moved on the m/z axis accordingly leaving the unannotated peaks in place.

The generated decoy spectra are concatenated to the original spectra in the library, and the search is run against this concatenated target-decoy spectral library. The corrected FDR value is then calculated as described previously in Sticker et al.[11].

$$FDR = \frac{\#decoy}{\#target}$$

Percolator rescoring

We integrated Percolator[13] version v3-04 into COSS[7] such that Percolator's input features are generated from COSS results and Percolator can be executed automatically. In addition to the main scoring function, MSROBIN, which is based on probabilistic scoring, COSS implements multiple scoring functions like Cosine similarity, MSE (Mean Square Error for both intensity and m/z values), Spearman correlation, Pearson correlation with and without using log transform in order to provide Percolator with more score features. To maximize the amount of information available for rescoring, additional features are calculated: MatchedPeaksQueryFraction (number of matched peaks divided by number of total query peak), MatchedPeaksLibFraction (number of matched peaks divided by number of total query peaks of a given libraryspectrum), SumMatchedIntQueryFraction (sum of intensities of matched peaks divided by sum of intensities of total query spectrum) and MatchedIntLibFraction (sum of intensities of matched peaks divided by sum of intensities of total peaks of library spectrum). In total, COSS provides 22 features to Percolator. All features with detailed descriptions are found in Supplementary Table S-2. For MSPepSearch, Percolator was provided with the full list of available features from the tool's output (10 features, Supplementary Table S-2). In the case of MS-GF+, the Percolator input text file is generated using the msgf2pin command provided by MS-GF+[15] (31 features).

Percolator was run with the target-decoy competition method enforced as all searches are performed against a concatenated database. In addition, the error check is overridden to ensure the output contains the rescored q-values as obtained from the Support Vector Machine (SVM). Identifications from the fractionated sample were concatenated prior to rescoring without filtering the result at 1%FDR.

Retention time prediction

In order to evaluate how rescoring with Percolator improves true positive identifications, we used an orthogonal validation using predicted retention time which makes use of DeepLC[26] , a deep learning algorithm that accurately predicts peptide retention times. In order to run DeepLC (version 0.1.35), we have first generated a comma separated file (CSV) having columns of PSM ID, sequence, modification and observed retention time from the search results using custom scripts. In addition, to obtain more accurate DeepLC prediction, DeepLC calibration files were constructed by selecting the 1000 peptides with the highest score across ten equally sized retention time windows.

Peak intensity prediction

The second orthogonal validation method we have used to evaluate the improvement of true positive results due to Percolator is the correlation of predicted intensity versus observed intensities for each PSM's. For this we have used MS2PIP[27–29], a machine learning tool capable of accurate prediction of peak intensities of a given peptide sequence. To run MS2PIP (v3.8.0), input files were generated as a tab delimited file with columns PSM ID, modification, peptide sequence and charge. In addition to this generated file, MS2PIP requires a configuration file and the MGF file of the corresponding dataset. MS2PIP is then executed with -x option to get the calculated pearson correlation between the observed peak intensities and the predicted peak intensities.

RESULTS AND DISCUSSION

Effects of different feature sets on the rescored result

Percolator uses a machine learning algorithm to rescore the output of peptide identification tools to achieve higher sensitivity and specificity. The result of this process depends strongly on the quality and utility of the input features. Hence, we have supplemented standard COSS output with additional scores and features for each PSM to maximize the potential of the rescoring step. Principal component analysis (Supplementary Figure S-2) shows several distinct clusters in the features, indicating that each group of features has the potential to add unique information to the Support Vector Machine (SVM) trained by Percolator. To examine the effect of these features on rescoring in spectral library searching, we analyzed different combinations of input features with Percolator. The full set of inputs to Percolator and their description can be found in Supplementary Table S-2. Figure 1 shows that using all features from COSS yields the best results in terms of identification rate at 1% FDR . While removal of precursor mass and charge has little effect on the identification rate, the intermediate scores generated by COSS (peak and intensity fraction of matched spectra) does lower the identification rate substantially . As these features are the main parameters to calculate the score, this is not unexpected. These differences in identification rate are consistent across a large q-value range (Supplementary Figure S-3).

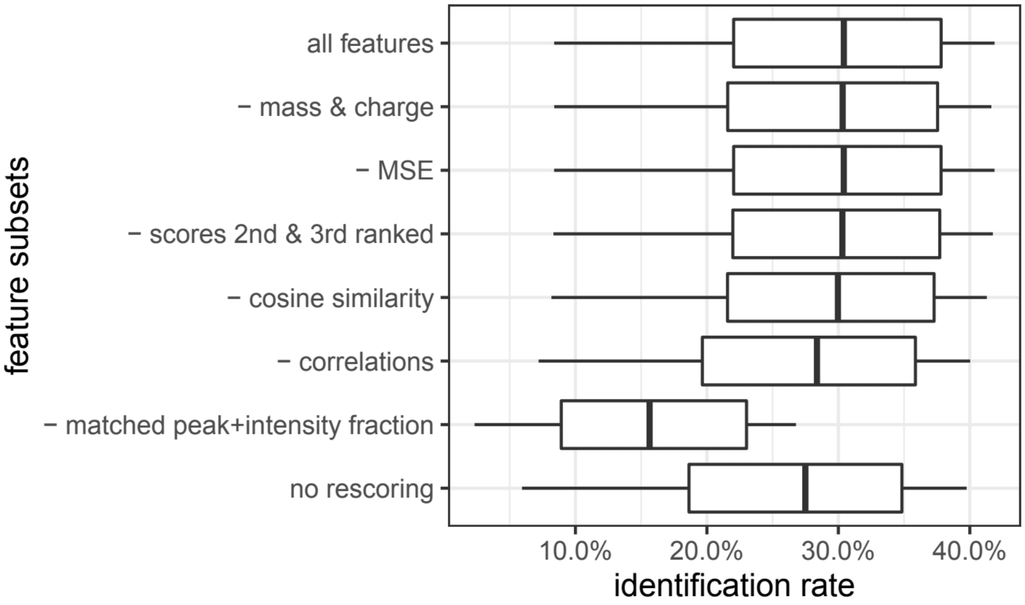

Figure 1. Effect of different feature sets provided to Percolator as measured by identification rate at 1% FDR.

Comparison of COSS, MSPepSearch and MS-GF+ with and without Percolator

To put the improvement observed in rescoring COSS results into context, we compared these with rescoring results from MSPepSearch, a different but performant spectral library search tool, and MS-GF+, a popular database search engine. Figure 2 shows the identification rates obtained either at 1% FDR (without Percolator), or for a q-value at or below 0.01 (with Percolator) for each of the two test data sets as analyzed by COSS, MSPepSearch, and MS-GF+. For all datasets, the COSS search against MassIVE consistently outperforms both MSPepSearch and MS-GF+, and this for both pre- and post-Percolator identification results. In the case of data sets from PXD010154, MS-GF+ has slightly more identifications than COSS against the NIST spectral library, which can likely be attributed to incomplete coverage of the library for this sample. This effect is also reflected in the even lower performance of MSPepSearch on this data set against the NIST library. In the case of PXD013477, identification is quite high for all three tools, in line with the results in the original manuscript[19]. Overall, these results show that rescoring the output of spectral library search engines can drastically increase identification rate. Depending on the coverage of the spectral library, the identification rate can exceed those obtained with database search engines.

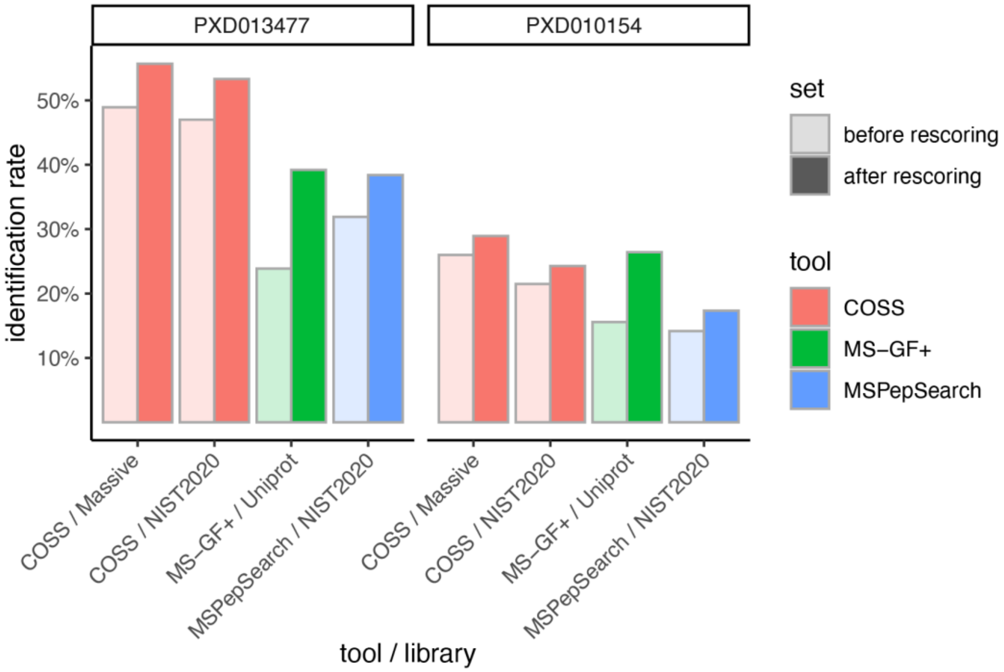

Figure 2. Comparison of identification rates achieved by COSS, MSPepSearch and MS-GF+, before rescoring and after rescoring. All results are either taken at 1% FDR (before rescoring) or at a q-value cut-off at 0.01 (after rescoring).

Comparison of obtained identifications before and after Percolator rescoring

To gain further insight into the difference between results obtained with, and without Percolator rescoring, the overlap at 1% FDR before and after rescoring was analyzed at the peptide level. Figure 3 and Figure S-4 show this identification agreement for COSS with and without Percolator at peptide and protein level respectively. For all data sets, Percolator removes some of the identified peptides (PXD013477: 3%; PXD010154: 3.4%), while adding a larger set of new results (PXD013477: 16%; PXD010154: 13%). The vast majority (PXD013477: 97%; PXD010154: 96.6%) of the identifications is maintained, however. Upon comparing the length of the peptides added and removed by Percolator, it appears that Percolator improves the sensitivity for small peptides (median length of ten amino acids), while removing preferentially larger peptides (Figure S-5). No significant difference in the amino acid composition of added *versus* removed peptides can be observed (Figure S-6). When comparing the number of peaks for each peptides added or removed by percolator (Figure S-7), only minor differences can be observed that are inconsistent across datasets. Finally, we have also compared the charge state of added and removed peptides and found that rescoring consistently affected low charge states more than higher (Figure S-8).

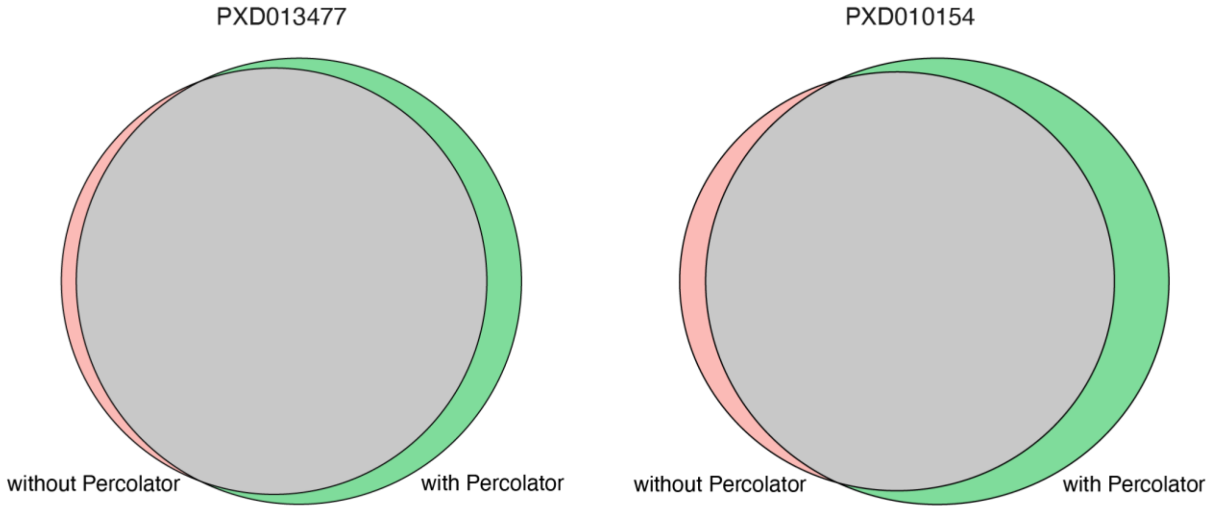

Figure 3. Peptide level overlap in search results from COSS against the NIST library at the 1% FDR level before and after Percolator rescoring. The number of peptides added by rescoring exceeds the number of removed peptides, resulting in an overall increase of the identification rate.

Orthogonal validation of the rescoring results using measured and predicted retention time

With the drastically increased reliability of retention time prediction, the comparison of the observed retention time and the predicted retention time for a specific peptide has been put forward as a means of orthogonal validation for PSMs[30,31]. Here, an overall poor correlation of predicted and observed retention time is indicative of a misidentified peptide. We have performed retention time prediction using DeepLC and compared the predicted with the observed retention time for PSMs affected by Percolator (Figure 4 and Figure S-9). The identifications added by rescoring had an overall smaller absolute difference in predicted *versus* observed retention time than the identifications removed by rescoring. This indicates that the latter set contains more misidentifications that are correctly removed by Percolator. In addition, the distribution of the difference in predicted *versus* observed retention time of Percolator added identifications closely resembles the distribution of the unchanged identifications, suggesting that these are indeed valid identifications.

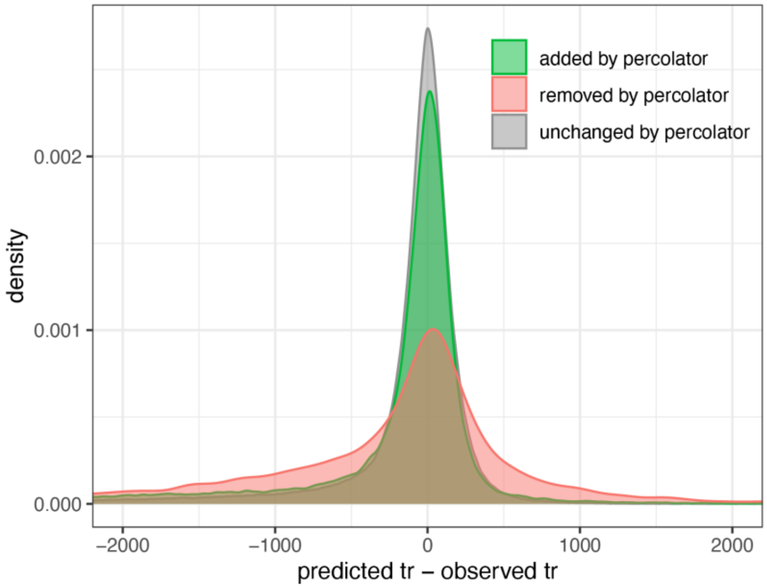

Figure 4. Comparison of measured and predicted retention time of peptide identification added and removed by rescoring. Shown here are identifications obtained with COSS on the PXD010154 data set and the NIST spectral library. The high resemblance between the distribution for the added and unchanged identifications suggests that there are indeed valid PSMs while the discrepancy with the distribution of the removed identifications indicates that those are enriched with misidentifications.

Orthogonal validation using measured and predicted peak intensity

In addition to retention time, we have used predicted peak intensities as another means of orthogonal validation for the increase of true positive identifications by Percolator. MS2PIP[27–29], a machine learning tool capable of accurate prediction of MS2 peak intensities, is used for peak intensity prediction. We have analyzed and compared the Pearson correlation between observed and predicted peak intensities of removed peptides, added peptides and peptides that are not affected by Percolator. The distributions of these correlation for peptides that are added and those that remained unchanged due to rescoring follow a similar pattern (Figure 5). On the other hand, the distribution for deleted peptides is very different witch an apex close to 0, which again indicates that deleted peptides are the result of misidentifications that are removed by Percolator. This result is consistent across all analyzed datasets and spectral libraries (Figure S-10)

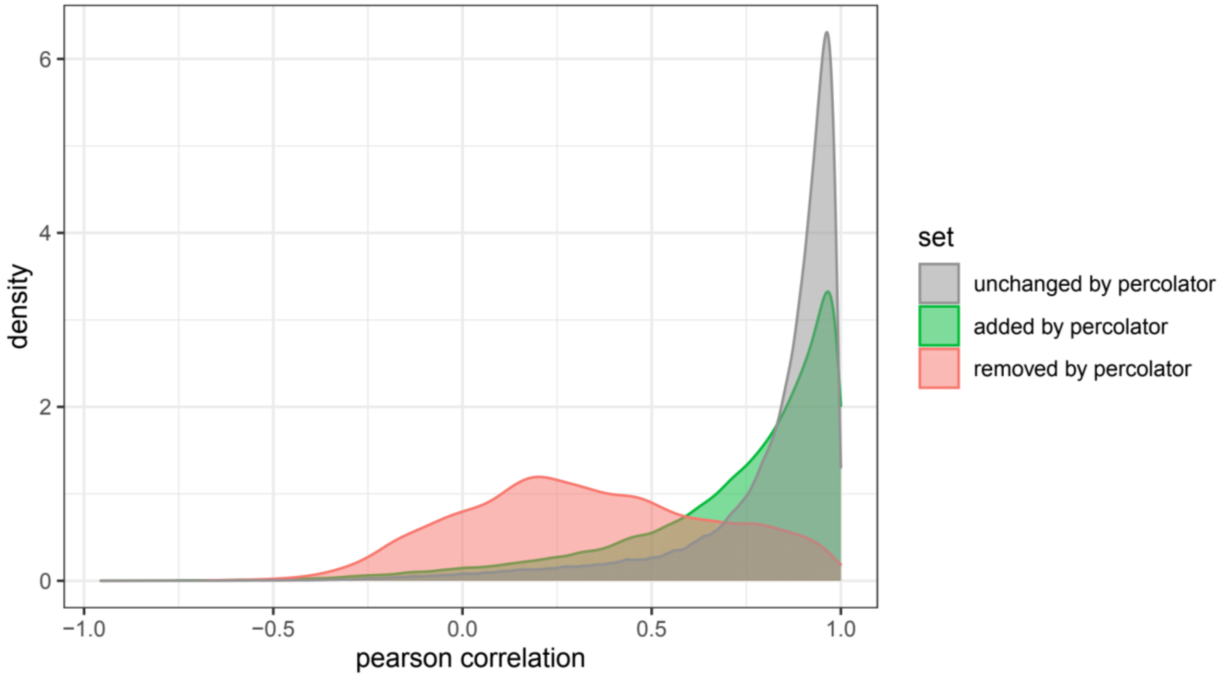

Figure 4. Comparison of measured and predicted peak intensity of peptide sequence added and removed by rescoring. The result is shown for the PXD010154 data set result of COSS against the NIST spectral library.

CONCLUSION

Increasing sensitivity while maintaining specificity is key in peptide and protein identification from mass spectrometry data. Rescoring peptide identifications using tools such as Percolator to increase sensitivity and specificity is thus common practice in sequence database searching. However, until now it was unclear if such post-processing had any benefits for spectral library searching. Here, we have shown that combining Percolator with such tools enhances sensitivity, and that it can also enhance specificity. Specifically, our COSS spectral library search tool shows increased sensitivity, and dramatically enhanced specificity when combined with Percolator due to additional features provided from COSS output. In addition to the main scoring function, we have implemented different scoring functions and their corresponding scoring values are included as features for Percolator. And we have shown how these additional features benefit the final Percolator results. We have thus shown that, for COSS at least, the combination of these benefits justifies the added complexity of the percolator post-processing step. Based on these findings, we can thus recommend use of rescoring in spectral library searching. To this end, the latest version of COSS (COSS-2.0) has been fitted with Percolator integration.

Supporting information:

The following supporting information is available free of charge at ACS website
http://pubs.acs.org

Figure1: Screen shot showing what settings we have used to run MsConvert to generate mgf files from RAW files. Figure 2: Principal component analysis of the different features generated by COSS as an input for Percolator. Figure 3: Identification rate of Percolator rescoring for different feature combination at different q-values. Figure 4: Protein level overlap in search results from COSS against the NIST library at the 1% FDR level before and after Percolator rescoring. Figure 5: Peptide length differences for peptides added or removed due to rescoring. Figure 6: Amino acid composition differences for peptides added or removed due to rescoring. Figure 7: Peptide peak count differences for those added or removed due to rescoring. Figure 8: Charge state difference for peptides added or removed due to rescoring: (a) dataset PXD010154 against NIST2020, (b) dataset PXD010154 against MassIVE library, (c) dataset PXD013477 against NIST2020 and (d) dataset PXD013477 against MassIVE library. Figure 9: Comparison of observed and predicted retention time fo added and removed peptides after rescoring using Percolator for: (a) dataset PXD010154 against MassIVE library, (b) dataset PXD013477 against NIST20 library and (c) dataset PXD013477 against MassIVE library. Figure 10: Comparison of observed and predicted peak intensity of added and removed peptides after rescoring using Percolator for: (a) dataset PXD010154 against MassIVE library, (b) dataset PXD013477 against NIST20 library and (c) dataset PXD013477 against MassIVE library. Table 1: Benchmarking data sets used to test COSS. The first four data sets are taken from “Assessing the relationship between mass window width and retention time scheduling on protein coverage for data-independent acquisition” (ProteomeXchange ID PXD013477), and the last thirty-six data sets are taken from the deep proteome and transcriptome abundance atlas (ProteomeXchange ID PXD010154). Individual raw files are selected randomly from their respective tissue. Table 2: Spectral libraries used for spectral library searching tools. Table 3: Percolator input features used for COSS and MsPepSearch. These features are collected from the output of each tool.

AVAILABILITY

The COSS software and its source code can be freely downloaded from https://github.com/compomics/COSS and is licensed under the permissive, open-source Apache License, version 2.0.

We have also provided configuration files, sample files, and our scripts to generate input files for DeepLC and MS2PIP on https://github.com/compomics/COSS-Percolator-manuscript.

ACKNOWLEDGMENTS

This project is supported by the National Institute of Health (NIH) [NCI-ITCR grant number 1U24CA199347 to G.A.S.], Research Foundation - Flanders (FWO) [grant number 1S50918N to R.G., grant number 1S90918N to T.V.D.B., G042518N and G028821N to L.M., and grant number 1253321N to P.V.], by the European Union's Horizon 2020 Program (H2020-INFRAIA-2018-1) [823839 to L.M.], and Ghent University Concerted Research Action [grant number BOF21-GOA-033 to L.M.]. R.B. acknowledges funding from the Vlaams Agentschap Innoveren en Ondernemen under project number HBC.2020.2205. The authors would like to thank all CompOmics group members for their ideas, discussions, and support.

REFERENCES

(1) 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. **2020**, No. July 2017, 292–306. https://doi.org/10.1002/mas.21543.

(2) Lam, H.; Aebersold, R. Using Spectral Libraries for Peptide Identification from Tandem Mass Spectrometry (MS/MS) Data. *Current Protocols in Protein Science* **2010**, *2010*. https://doi.org/10.1002/0471140864.ps2505s60.

(3) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. *Journal of Proteome Research* **2010**, *9* (1), 605–610. https://doi.org/10.1021/pr900947u.

(4) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *Journal of the American Society for Mass Spectrometry* **1994**. https://doi.org/10.1016/1044-0305(94)87009-8.

(5) Bittremieux, W.; Meysman, P.; Sta, W.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *Journal of Proteome Research* **2018**, *17*, 3463–3474. https://doi.org/10.1021/acs.jproteome.8b00359.

(6) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification Research Articles. **2006**, 1843–1849. https://doi.org/10.1021/pr0602085.

(7) Shiferaw, G. A.; Vandermarliere, E.; Hulstaert, N.; Gabriels, R.; Martens, L.; Volders, P. COSS: A Fast and User-Friendly Tool for Spectral Library Searching. *Journal of Proteome Research* **2020**, *19*. https://doi.org/10.1021/acs.jproteome.9b00743.

(8) Zhang, X.; Li, Y.; Shao, W.; Lam, H. Understanding the Improved Sensitivity of Spectral Library Searching over Sequence Database Searching in Proteomics Data Analysis. *PROTEOMICS* **2011**, *11* (6), 1075–1085. https://doi.org/10.1002/pmic.201000492.

(9) Ahrné, E.; Masselot, A.; Binz, P. A.; Müller, M.; Lisacek, F. A Simple Workflow to Increase MS2 Identification Rate by Subsequent Spectral Library Search. *Proteomics* **2009**, *9* (6), 1731–1736. https://doi.org/10.1002/pmic.200800410.

(10) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667. https://doi.org/10.1002/pmic.200600625.

(11) 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.

(12) Benjamini, Y. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. **2014**, No. November 1995. https://doi.org/10.2307/2346101.

(13) 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* (11), 923–925. https://doi.org/10.1038/nmeth1113.

(14) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *Journal of Proteome Research* **2009**, *8*, 3176–3181. https://doi.org/10.1021/pr800982s.

(15) Fast and Accurate Database Searches with MS-GF+Percolator. **2008**, No. July.

(16) Silva, A. S. C.; Bouwmeester, R.; Martens, L.; Degroeve, S. Data and Text Mining Accurate Peptide Fragmentation Predictions Allow Data Driven Approaches to Replace and Improve upon Proteomics Search Engine Scoring Functions. **2019**, *35* (May), 5243–5248. https://doi.org/10.1093/bioinformatics/btz383.

(17) Wilhelm, M.; Zolg, D. P.; Graber, M.; Gessulat, S.; Schmidt, T.; Schnatbaum, K.; Schwencke-Westphal, C.; Seifert, P.; de Andrade Krätzig, N.; Zerweck, J.; Knaute, T.; Bräunlein, E.; Samaras, P.; Lautenbacher, L.; Klaeger, S.; Wenschuh, H.; Rad, R.; Delanghe, B.; Huhmer, A.; Carr, S. A.; Clauser, K. R.; Krackhardt, A. M.; Reimer, U.; Kuster, B. Deep Learning Boosts Sensitivity of Mass Spectrometry-Based Immunopeptidomics. *Nature Communications* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-23713-9.

(18) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; Frejno, M.; Schmidt, T.; Schnatbaum, K.; Wilhelm, M.; Ponten, F.; Uhlen, M.; Gagneur, J.; Hahne, H.; Kuster, B. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Molecular system biology* **2019**, 1–16. https://doi.org/10.15252/msb.20188503.

(19) Li, W.; Chi, H.; Salovska, B.; Wu, C.; Sun, L.; Rosenberger, G.; Liu, Y. Assessing the Relationship Between Mass Window Width and Retention Time Scheduling on Protein Coverage for Data-Independent Acquisition. *Journal of the American Society for Mass Spectrometry* **2019**, *30* (8), 1396–1405. https://doi.org/10.1007/s13361-019-02243-1.

(20) Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Frewen, B.; Baker, T. A.; Brusniak, M.; Paulse, C.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L. HHS Public Access. *Nature Biotechnology* **2013**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.A.

(21) Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* **2018**, *7* (4), 412-421.e5. https://doi.org/10.1016/j.cels.2018.08.004.

(22) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nature Communications* **2014**, *5*, 5277.

(23) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *Journal of Proteome Research* **2018**, *17* (7), 2552–2555. https://doi.org/10.1021/acs.jproteome.8b00175.

(24) The Uniprot Consortium. UniProt : A Worldwide Hub of Protein Knowledge. *Nucleic Acids Research* **2019**, *47* (November 2018), 506–515. https://doi.org/10.1093/nar/gky1049.

(25) Zhang, Z.; Burke, M.; Mirokhin, Y. A.; Tchekhovskoi, D. v.; Markey, S. P.; Yu, W.; Chaerkady, R.; Hess, S.; Stein, S. E. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *Journal of Proteome Research* **2018**, *17* (2), 846–857. https://doi.org/10.1021/acs.jproteome.7b00614.

(26) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. *DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications*; 2020. https://doi.org/10.1101/2020.03.28.013003.

(27) Degroeve, S.; Maddelein, D.; Martens, L. MS2PIP Prediction Server: Compute and Visualize MS2 Peak Intensity Predictions for CID and HCD Fragmentation. *Nucleic Acids Research* **2015**, *43* (W1), W326–W330. https://doi.org/10.1093/nar/gkv542.

(28) Gabriels, R.; Martens, L.; Degroeve, S. Updated MS^{2}PIP Web Server Delivers Fast and Accurate MS^2 Peak Intensity Prediction for Multiple Fragmentation Methods, Instruments and Labeling Techniques. *Nucleic Acids Research* **2019**, *47* (W1), W295–W299. https://doi.org/10.1093/nar/gkz299.

(29) Degroeve, S.; Martens, L.; Jurisica, I. MS2PIP: A Tool for MS/MS Peak Intensity Prediction. *Bioinformatics* **2013**, *29* (24), 3199–3203. https://doi.org/10.1093/bioinformatics/btt544.

(30) Gong, S.; Gaccioli, F.; Dopierala, J.; Sovio, U.; Cook, E.; Volders, P. J.; Martens, L.; Kirk, P. D. W.; Richardson, S.; Smith, G. C. S.; Charnock-Jones, D. S. The RNA Landscape of the Human Placenta in Health and Disease. *Nature Communications* **2021**, *12* (1), 1–17. https://doi.org/10.1038/s41467-021-22695-y.

(31) Ruiz Cuevas, M. V.; Hardy, M. P.; Hollý, J.; Bonneil, É.; Durette, C.; Courcelles, M.; Lanoix, J.; Côté, C.; Staudt, L. M.; Lemieux, S.; Thibault, P.; Perreault, C.; Yewdell, J. W. Most Non-Canonical Proteins Uniquely Populate the Proteome or Immunopeptidome. *Cell Reports* **2021**, *34* (10). https://doi.org/10.1016/j.celrep.2021.108815.