UC San Diego

UC San Diego Previously Published Works

Title

Binary Classifier for Computing Posterior Error Probabilities in MetaMorpheus

Permalink

https://escholarship.org/uc/item/68g1d8cb

Journal

Journal of Proteome Research, 20(4)

ISSN

1535-3893

Authors

Shortreed, Michael R
Millikin, Robert J
Liu, Lei
et al.

Publication Date

2021-04-02

DOI

10.1021/acs.jproteome.0c00838

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2022 April 02.

Published in final edited form as:

J Proteome Res. 2021 April 02; 20(4): 1997–2004. doi:10.1021/acs.jproteome.0c00838.

A Binary Classifier for Computing Posterior Error Probabilities in MetaMorpheus

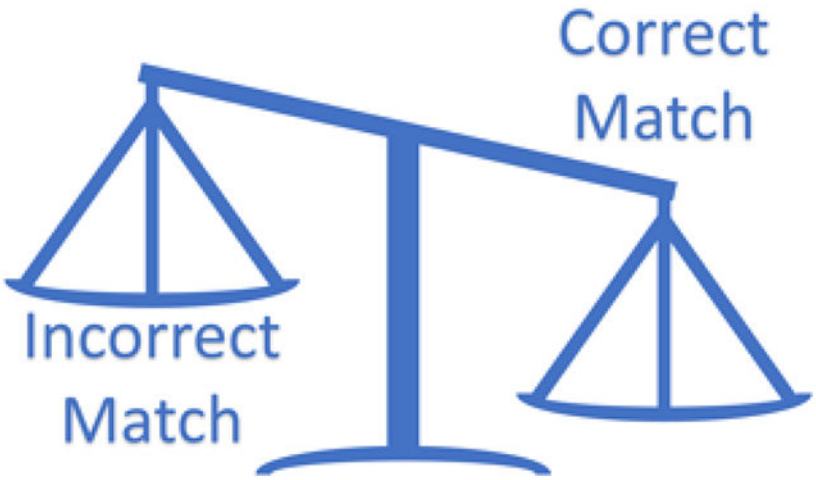
Michael R. Shortreed[1], **Robert J. Millikin**[1], **Lei Liu**[1], **Zach Rolfs**[1], **Rachel M. Miller**[1], **Leah V. Schaffer**[1], **Brian L. Frey**[1], **Lloyd M. Smith**[1,*]

[1]Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA

Abstract

MetaMorpheus is a free, open-source software program for the identification of peptides and proteoforms from data-dependent acquisition, tandem MS experiments. There is inherent uncertainty in these assignments for several reasons including: limited overlap between experimental and theoretical peaks; m/z uncertainty; and noise peaks or peaks from co-isolated peptides that produce false matches. False discovery rates provide only a set-wise approximation for incorrect spectrum matches. Here, we implemented a binary decision tree calculation within MetaMorpheus to compute a posterior error probability, which provides a measure of uncertainty for each peptide spectrum match. We demonstrate its utility for increasing identifications and resolving ambiguities in bottom-up, top-down, proteogenomic and non-specific digestion searches.

Graphical Abstract



Keywords

MetaMorpheus; proteomics; DDA; binary decision tree; search engine; bottom-up; top-down; posterior error probability; proteogenomics; open-source

*Corresponding author. Tel: 1-608-263-2594. smith@chem.wisc.edu.

Supporting Information

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Introduction

The Morpheus search algorithm[1] was originally created in 2013 to accommodate an increased prevalence of high-resolution tandem mass spectra (MS/MS) in proteomics. The algorithm took advantage of the specificity provided by high mass accuracy to assign charge states and remove non-monoisotopic peaks, but with minimal loss of sensitivity. The scoring algorithm considered only the number of matching products plus the fraction of spectrum abundance assigned to matching products. This modest program, remarkable in its simplicity, yielded excellent results.

Our group developed an interest in identifying a multiplicity of post-translational modifications (PTMs) within a single search on data acquired from unenriched samples. At that time, identifying PTMs was performed primarily on samples where the PTM of interest was enriched and the modification was set as variable within the search engine. This strategy has been used conventionally for many years, but the variable modification strategy fails to yield results with high confidence when the number of modified peptides represents a small fraction of the total. Our idea was to allow variable modifications strategically only at annotated positions within the proteome and nowhere else. This worked remarkably well, permitting the analysis of dozens of different PTM types within a single search and yielding identifications with high confidence. We joined forces with the Morpheus team and released an updated version with this new capability.[2] Subsequently, we extended this work to cover previously unannotated modifications through a two-pass search algorithm dubbed global post-translational modification discovery (G-PTM-D).[3]

These early successes spawned many new ideas and eventually the need to release our own software program, MetaMorpheus[4], to accommodate the growing functionality. MetaMorpheus now has capacity for mass calibration, label-free quantification[5], top-down search, crosslink search[6], discovery of O-glycosylated peptides[7] and non-specific searches[8]. One can also conduct a single search with multiple proteases[9], improving protein inference over single-protease approaches. However, the scoring algorithm, until recently, had evolved little and the only statistical metric provided was a group-wise false discovery rate (FDR) reported in MetaMorpheus as a q-value[10–11]. One important value that was greatly needed was an individual confidence measure for each peptide spectrum match (PSM) or proteoform spectrum match (PrSM), peptide or proteoform identification. This information is valuable as one begins the process of validating and interpreting proteomics results. Early approaches to this were reported by Keller[12] and then by Anderson[13]. To obtain individual confidence metrics for MetaMorpheus identifications, one could manually calculate a posterior error probability (PEP) by determining the local FDR[11] for each set of matches with the same MetaMorpheus score. Or one could post-process results using software created by other groups (e.g., Percolator[14–18] or Peptide Prophet[12]).

Here we implement a binary decision tree[19] (BDT) in MetaMorpheus that computes the PEP for each spectrum match. The PEP of an individual PSM represents the probability that the identification is incorrect. The PEP is effectively an optimized scoring metric when arrived at using the BDT algorithm. This optimization allows greater discrimination between correct

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and incorrect matches. The essence of a BDT is to ask a series of true/false questions, one for each attribute considered, to assign each candidate to one of two groups (see Figure 1). In this work, the candidates are spectrum matches and the attributes include the fraction of matched intensity, the longest uninterrupted sequence of matched fragment ions, the number of missed cleavages, etc. (see Table 1). The two groups are correct and incorrect matches. Each question can be asked only once along a path and the order the questions are asked is optimized automatically for efficiency and accuracy. The BDT is trained on a subset (75%) of spectrum matches from a target/decoy search and then applied to and validated on the remaining 25% of spectrum matches. This process is repeated four times so that no spectrum match is scored using a training set in which it was included.

A major advantage of the BDT is that a new model can be quickly generated for each search. MetaMorpheus is a flexible search engine capable analyzing many different datatypes using any number of proteases, fragmentation types/energies, instrument resolutions, and other parameters. Therefore, it needed an approach to computing spectrum match PEPs with comparable flexibility. BDTs can be trained rapidly using a wide variety of different attributes. Below, we describe our implementation of the BDT in MetaMorpheus and provide results for searches of bottom-up, top-down, non-specific and proteogenomic data.

Experimental

The MetaMorpheus search software, which includes the BDT functionality, is coded in the C# programming language. This software is open source and freely available with a permissive MIT license (https://github.com/smith-chem-wisc/MetaMorpheus). MetaMorpheus is also available as a Docker container (https://hub.docker.com/r/smithchemwisc/metamorpheus). The MetaMorpheus Windows Graphical User Interface (GUI) requires a 64-bit operating system and .NET Core 3.1. The command-line version of MetaMorpheus supports any operating system that supports .NET Core including Windows, MacOS and Linux. MetaMorpheus supports parallelization, using n-1 available logical processors by default. Users are free to select the number of logical processors used. A minimum of 8GB of RAM is recommended but higher amounts of RAM will speed up performance. A simple search of a conventional bottom-up run with a single processor can be finished in a matter of a few minutes. New users are encouraged to test installation of MetaMorpheus using a variety of test data sets available with instructions on the MetaMorpheus GitHub page. An extensive Wiki is also provided there, that covers typical usage and a glossary of terminology. Users with questions or experiencing problems can contact us via the Issues tab of the GitHub page or at our email address (mm_support@chem.wisc.edu). MetaMorpheus uses a FastTree Binary Classifier (https://www.nuget.org/packages/Microsoft.ML.FastTree/1.3.1) included via Nuget package. FastTree's binary classification boosting framework's natural probabilistic interpretation is explained in "From RankNet to LambdaRank to LambdaMART: An Overview" by Chris Burges(https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/MSR-TR-2010-82.pdf). All analyses were performed on a computer running Microsoft Windows 10.0.19041 with a 64-bit Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz processor with 28 threads and 128GB installed RAM.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results and Discussion

Binary search.

Binary classifiers divide a collection of objects into two groups (e.g. True and False). This is accomplished by applying a series of challenges that maximize separation between the groups. Each challenge is applied only once along the branch of the tree. The order of challenges is chosen automatically by the algorithm to maximize group separation at each step. Here, group one is confident target matches, while group two is decoy matches. The distribution of the two groups in the leaves of the tree provide the group assignment probability. In this work, the fraction of group two matches in each of the leaves provides the probability that any member placed in that leaf by application of the model is an incorrect match. (Figure 1)

Parameters.

Fourteen different attributes (Table 1) are used in the process. A list of attributes and their definitions are provided in Table 1. Plots (Figure 2) show how the fraction of true positive target PSMs to total PSMs varies across the range of respective values observed in one bottom-up search (*see* Bottom-Up Vignette *below*). The fraction of true positive target matches varies for the fourteen different attributes. The fraction varies strongly for Total Matching Fragment Count, Intensity, PSM Count, Complementary Ion Count and Delta Score. The fraction varies weakly, but measurably, for the remaining attributes.

Construction of the model and placement within the workflow.

Creation and application of the BDT for computing the Posterior Error Probability (PEP) for each PSM occurs after the search is completed and all PSMs have been assigned. Each PSM is assigned a separate q-value depending on its MetaMorpheus score rank. The q-value is the false discovery rate for all PSMs with MetaMorpheus score at or above a specified score threshold. This q-value is used to select members of the training sets (see below). We compute a PEP-derived q-value[16], referred to hereafter as the PEP q-value, after training is completed and the models have been applied to the PSMs. The PEP q-value is the average of all individual PEP values for a group of PSMs or peptides down to and including the current PSM or peptide, after sorting all PSMs and peptides in descending order by PEP.[20] The PEP q-value is a comparable metric to the traditional q-value.

Training and testing.

We employed a cross-validation approach[21] for training and testing of the model. Training and testing sets of PSMs are chosen randomly (each training set contains 75% of the total PSMs, up to 1 million). The trained model is then applied to and evaluated on the remaining 25% of PSMs. This process is repeated a total of four times such that no PSM is evaluated using a model that included it with the training. Target hits with q-value < 0.01 are used as correct matches, and decoy hits are used as incorrect matches. There is no overlap between members of the training and testing sets. Training occurs in a single round with 400 trees in the ensemble, which is the default for the FastTree Binary Classifier.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Model Performance Metrics.

The model developed during the training phase is applied to the test set. The performance of the model is reported using several different metrics (Table 2.) The Count of Ambiguous Peptides Removed is a special feature of MetaMorpheus requiring further explanation. The MetaMorpheus search compares all spectra against all theoretical peptides/proteoforms whose intact masses agree within some specified tolerance (e.g., 10 ppm). The MetaMorpheus score for each potential match is computed. The integer value of the MetaMorpheus score is the count of matched fragment ions. The decimal portion of the MetaMorpheus score is the spectrum intensity fraction accounted for by the matched fragment ions. It is not unusual for multiple unique theoretical peptides to have the same MetaMorpheus score (+/− 1E-09) for a single spectrum. We refer to this as an ambiguous PSM and we report all theoretical peptide sequences in the output for the PSM separated by the "|" character. PSM ambiguities can arise from target or decoy peptides. We use the BDT model to resolve many of these ambiguities. A separate PEP is computed for each peptide possibility in the ambiguous assignment. Whenever the PEP for a peptide possibility in an ambiguous assignment is at least 5% lower than the other possibilities, that ambiguous assignment is chosen as the most likely assignment and the other possibilities are removed. Thus, while the Count of Ambiguous Peptides Removed is a metric of the model, the actual resolving of ambiguities is a valuable feature of the BDT classification. Several example vignettes are described below. Numerical performance metrics for these examples are collectively reported in Table 3.

Bottom up vignette.

The experimental procedure for generation of the data set for the bottom-up vignette was reported previously.[9] Data was derived from a trypsin digest of 10^7 human Jurkat cells. Peptides were fractionated off-line by high-pH reverse-phase liquid chromatography prior to the LC–MS/MS analysis on a nanoACQUITY LC system (Waters, Milford, MA) interfaced with a Thermo Scientific LTQ Orbitrap Velos mass spectrometer. All mass spectrometry raw files are freely available on the MassIVE platform (https://massive.ucsd.edu; ID: MSV000083304; Files; 12-18-17_fract1-10).

The data analysis was performed using MetaMorpheus version 0.0.313. The following search settings were used: protease = trypsin; maximum missed cleavages = 2; minimum peptide length = 7; maximum peptide length = unspecified; initiator methionine behavior = Variable; fixed modifications = Carbamidomethyl on C, Carbamidomethyl on U; variable modifications = Oxidation on M; max mods per peptide = 2; max modification isoforms = 1024; precursor mass tolerance = ±5.0000 PPM; product mass tolerance = ±20.0000 PPM; report PSM ambiguity = True. The combined search database contained 20379 non-decoy protein entries including 0 contaminant sequences. The database was obtained in XML format from UniProt, downloaded 01/12/2021 and contained annotated PTMs, which are automatically detected with MetaMorpheus. The total time to perform the Search task on 10 spectra file(s) was 9.0 minutes. The time to perform the BDT analysis was 67 s. The final search tallies were 88484 target PSMS and 32621 peptides at q-value < 0.01. PEP q-values were then computed. The search yielded 92802 PSMs and 34506 peptides at PEP q-value

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

< 0.01, increases of 4318 (4.9%) and 1885 (5.8%) respectively. A total of 210 ambiguous peptides were disambiguated.

A second analysis was performed using an in-house created entrapment database (20379 protein entries) in addition to the human database. This entrapment database was created by fixing the position of lysine and arginine residues and randomizing the remaining amino acids on a protein-by-protein basis. The N-terminal methionine was also preserved when present. Annotated PTMs found in the original database were shifted to new positions along with their corresponding amino acid. Using a randomized version of the target database for entrapment is preferred over use of a database for another organism[22]. This entrapment analysis was used to evaluate performance as all PSMs assigned to entrapment peptides are presumably false positives. The second search yielded 85180 PSMs, including 380 false positive entrapment PSMs (0.45%) and 31426 peptides, including 155 false positive entrapment peptides (0.49%). After performing the BDT analysis, these values changed to 90524 PSMs, including 294 false positive entrapment PSMs (0.32%) and 33668 peptides, including 185 false positive entrapment peptides (0.55%). Please note that evaluations here and in vignettes below were performed using a single entrapment database. Therefore, the results do not represent the average results that would have been obtained had we repeated the experiment 10 or more times, each using a separately crafted and unique entrapment database.

HLA vignette.

The following vignette demonstrates the application of the BDT to peptides identified in a non-specific search of human HLA peptides, obtained from a study performed by Bassani-Sternberg and colleagues[23]. The data set used here can be obtained from the PRIDE repository using the identifier, PXD004894. The data files used here include 20141208- and 20141210_QEp7_MiBa_SA_HLA-I-p_MM15 samples 1-4, A & B (14 files). The following search settings were used: protease = non-specific; maximum missed cleavages = 19; minimum peptide length = 8; maximum peptide length = 20; initiator methionine behavior = Variable; fixed modifications = Carbamidomethyl on C, Carbamidomethyl on U; variable modifications = Oxidation on M; max mods per peptide = 2; max modification isoforms = 1024; precursor mass tolerance = ±6.0000 PPM; product mass tolerance = ±20.0000 PPM; report PSM ambiguity = True. The human search database contained 20379 non-decoy protein entries, downloaded from UniProt on 01/08/2021 in FASTA format, including 0 contaminant sequences.

The total time to perform the Search task on 14 spectra file(s) was 219.52 minutes. The time to perform the BDT analysis was 186 s. The final search tallies were 138450 target PSMS, 17789 peptides at q-value < 0.01. PEP q-values were then computed. The searched yielded 127958 PSMs and 21313 peptides at PEP q-value < 0.01, a decrease of 7.6% and increase of 19.8% respectively. A total of 115 ambiguous peptides were disambiguated.

In this non-specific search, the count of PSMs at 1% FDR decreased upon using the BDT while the number of unique peptides increased. There are a number of possible explanations for this behavior. In a non-specific search, the search space of theoretical peptides in both the forward target database and in the reverse decoy database are very high compared to a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

typical search with specific proteolytic cleavage sites. This significantly lowers sensitivity, which we observe as a high cut-off MetaMorpheus score at 1% peptide FDR. Such large databases are also often prone to a high false positive rate. We hypothesize that several medium- to low- scoring PSMs are in fact false positives which the BDT filters out; the BDT takes into consideration many additional facets of the PSM compared to simply ranking by the MetaMorpheus score. In contrast, for the peptides, which all here have a high MetaMorpheus score and so presumably are not false positives, are not filtered out by the BDT algorithm.

As above, a second analysis was performed using an entrapment database constructed similarly to the aforementioned entrapment database. However, no PTMs were included in the search. Therefore, no PTMs were included in the entrapment database. The second search yielded 117109 PSMs, including 584 false positive entrapment PSMs (0.50%) and 15090 peptides, including 73 false positive entrapped peptides (0.48%). After performing the BDT analysis, these values changed to 100433 PSMs, including 193 false positive entrapment PSMs (0.19%) and 17459 peptides, including 72 false positive entrapment peptides (0.41%). The BDT analysis decreased the number of identifications by 16676 PSMs (14.2%) and increased the number of 2369 peptides (15.7%), while reporting fewer entrapped false positive PSMs and a similar number of entrapped false positive peptides.

Top-down vignette.

Data for the top-down vignette are from a study of mouse mitochondria.[24] All mass spectrometry raw files are freely available on the MassIVE platform (https://massive.ucsd.edu; ID: MSV000082366). The files included 08-02 and 08-03-17_B9_myoblast_A fractions 1-12, reps 1 and 2 (12 files). The data analysis was performed using MetaMorpheus version 0.0.313.

The following search settings were used: protease = top-down; maximum missed cleavages = 2; minimum peptide length = 7; maximum peptide length = unspecified; initiator methionine behavior = Variable; fixed modifications = 0; variable modifications = 0; max mods per peptide = 2; max modification isoforms = 1024; precursor mass tolerance = ±10.0000 PPM; product mass tolerance = ±20.0000 PPM; report PSM ambiguity = True. The mouse search database, downloaded from UniProt in XML format on 01/12/2021, contained 17051 non-decoy protein entries including 0 contaminant sequences. The total time to perform the Search task on 12 spectra file(s) was 14.7 minutes.

The time to perform the BDT analysis was 20 s. The final search tallies were 11365 target PrSMs and 809 proteoforms at q-value < 0.01. PEP q-values were then computed. The search yielded 11724 PrSMs and 873 proteoforms at PEP q-value < 0.01, increases of 359 (3.2%) and 64 (7.9%) respectively. A total of 506 ambiguous proteoforms were disambiguated.

As above, a second analysis was performed using an in-house created entrapment database constructed in a similar fashion to the human entrapment database except using the mouse protein sequence database as input. The second search yielded 11077 PrSMs, including 35 false positive entrapment PrSMs (0.32%) and 808 proteoforms, including 1 false

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

positive entrapment proteoforms (0.12%). After performing the BDT analysis, these values increased to 11463 PrSMs, including 26 false positive entrapment PrSMs (0.23%) and 861 proteoforms, including 9 false positive entrapment proteoforms (1.05 %). The BDT analysis increased the number of identifications by 386 PrSMs (3.5%) and 53 proteoforms (6.6%), while reporting decreased entrapped PrSMs and an increase in entrapped proteoforms, although still approximately 1% false positives.

Proteogenomics Vignette.

The bottom-up vignette data was used for this analysis. A proteogenomic database was created with Spritz[25] (https://github.com/smith-chem-wisc/Spritz). Input for Spritz was obtained from www.ncbi.nlm.nih.gov using the following identifiers: SRR791578, SRR791579, SRR791580, SRR791581, SRR791582, SRR791583, SRR791584, SRR791585, SRR791586. Sequences were compared against the Ensembl Archive Release 82. Proteomics data analysis was performed using the Spritz-generated sample-specific proteogenomic database with MetaMorpheus version 0.0.313.

The Jurkat proteogenomic database was constructed using Spritz version 0.1.3. The paired-end RNA sequencing data used for database construction was previously obtained and accessed using GSE45428 in GEO SRA.[26] The workflow for database creation using Spritz has been described in detail.[27] In brief, genomic references including human genome and gene model files from Ensembl version 82 and known human variation sites are downloaded from dbSNP[28]. Next, skewer[29] is used to remove adapter sequences from RNA and filter out low quality reads. The reads are then aligned to the human reference genome using hisat2[30] before variant analysis is performed using the Genome Analysis Toolkit (GATK)[31–32] version 4.0.11.0. SnpEff[33] has been adapted to enable the annotation of discovered variants in Uniprot-XML formatted databases. Following variant annotation, post-translational modifications are transferred to the proteogenomic database from the human UniProt database (downloaded 6/30/2020).

The following search settings were used: protease = trypsin; maximum missed cleavages = 2; minimum peptide length = 7; maximum peptide length = unspecified; initiator methionine behavior = Variable; fixed modifications = Carbamidomethyl on C, Carbamidomethyl on U; variable modifications = Oxidation on M; max mods per peptide = 2; max modification isoforms = 1024; precursor mass tolerance = ±5.0000 PPM; product mass tolerance = ±20.0000 PPM; report PSM ambiguity = True. The combined search database contained 77534 non-decoy protein entries including 0 contaminant sequences. The total time to perform the Search task on 10 spectra file(s) was 18.72 minutes.

The time to perform the BDT analysis was 2.22 minutes. The final search tallies were 88849 target PSMS and 32671 peptides at q-value < 0.01. PEP q-values were then computed. The search yielded 93331 PSMs and 34777 peptides at PEP q-value < 0.01, increases of 8882 (5.0%) and 2106 (6.4%) respectively. A total of 786 ambiguous peptides were disambiguated. Because this is a proteogenomic search, we were interested in identifying peptides with amino acid variants. Here, we found 449 variant PSMs at q<0.01. After applying the BDT we found 455 variant PSMs at PEP q-value<0.01 with an overlap between

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the sets of 431. In terms of variant-containing peptides, we found 190 at q<0.01 and 193 after using the BDT at PEP q-value<0.01 with 183 peptides overlapping the two sets.

As above, a second analysis was performed using an additional human entrapment XML format database. The second search yielded 85584 PSMs, including 376 false positive entrapment PSMs (0.44%) and 31641 peptides, including 150 false positive entrapment peptides (0.47%). After performing the BDT analysis, these values increased to 91692 PSMs, including 219 false positive entrapment PSMs (0.24%) and 34181 peptides, including 132 false positive entrapment peptides (0.39%). Therefore, the BDT analysis increased the number of identifications by 6108 PSMs (7.1%) and 2540 peptides (8.0%), while reporting a decrease in the number of entrapped false positives.

Effect of individual components.

Example results from bottom-up, non-specific peptide and top-down searches can be seen in the vignettes above. The models constructed for the bottom-up and non-specific peptide searches used 13 attributes, skipping the variant attribute as no variant peptides were included in the database. We were interested in determining the effects of individual parameters on the complete model. This was accomplished by performing 13 separate bottom-up searches using 12 attributes and skipping the one under examination. A bar plot reporting the Accuracy for all 13 searches, labelled with the missing attribute, is shown in Figure 3A.

Next, we performed 12 more searches of the same data from the bottom-up vignette wherein we constructed the model one feature at a time, beginning with the attribute that had the highest impact (Delta Score). One attribute was added at a time in the order shown in the bar plot (Figure 3B). Each additional attribute improves the accuracy of the model.

Comparison to percolator.

The Percolator algorithm (v.3.0.4 http://percolator.ms/) is a support vector machine used to re-rank PSM and peptide identifications using user-supplied parameters. Percolator also reports peptide posterior error probabilities. It can perform a similar role to the BDT. The ability to perform the comparison was enabled by adding a new output to MetaMorpheus, designed specifically to meet the needs of Percolator input format. This Percolator input contains all the same features and values for each PSM and peptide used by BDT.

We repeat here the values observed for the bottom-up vignette search with the additional entrapment database for ease of readability. The search yielded 85180 PSMs, including 380 false positive entrapment PSMs (0.45%) and 31426 peptides, including 155 false positive entrapment peptides (0.49%). Percolator analysis of the results yielded 91593 PSMs including 380 false positive entrapment PSMs (0.41%) and 34715 peptides including 276 false positive entrapment peptides (0.80%). The BDT computation time was 68s and the Percolator computation time (using flags -U and --search-input concatenated) was 52s. These two results demonstrate that for this standard type of search, both the percolator algorithm and the BDT perform comparably well.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Conclusion.

The addition of a BDT to MetaMorpheus provides much needed statistical support for individual peptide and proteoform identifications. The computation time and performance are competitive with existing stand-alone programs such as percolator. In most cases, the BDT increases the number of PSM, peptide and proteoform identifications beyond the numbers reported using only the traditional q-value. In addition, it resolves many peptide assignment ambiguities that could not have been resolved using only the MetaMorpheus score. In future studies, we plan to integrate the computed peptide posterior error probabilities into MetaMorpheus' protein inference, which should further improve confidence in protein identification.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work was supported by NIH-NIGMS grant R35GM126914. RMM was supported in part by the NIH Chemistry–Biology Interface Training Grant (T32 GM008505). R.J.M. was supported by an NHGRI training grant to the Genomic Sciences Training Program 5T32HG002760. LVS was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number T32GM008349. We would also like to thank Austin V. Carr for his assistance in preparation of figures.

References

1. Wenger CD; Coon JJ, A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J Proteome Res2013, 12 (3), 1377–86. [PubMed: 23323968]
2. Shortreed MR; Wenger CD; Frey BL; Sheynkman GM; Scalf M; Keller MP; Attie AD; Smith LM, Global Identification of Protein Post-translational Modifications in a Single-Pass Database Search. J Proteome Res2015, 14 (11), 4714–20. [PubMed: 26418581]
3. Li Q; Shortreed MR; Wenger CD; Frey BL; Schaffer LV; Scalf M; Smith LM, Global Post-Translational Modification Discovery. J Proteome Res2017, 16 (4), 1383–1390. [PubMed: 28248113]
4. Solntsev SK; Shortreed MR; Frey BL; Smith LM, Enhanced Global Post-translational Modification Discovery with MetaMorpheus. J Proteome Res2018, 17 (5), 1844–1851. [PubMed: 29578715]
5. Millikin RJ; Solntsev SK; Shortreed MR; Smith LM, Ultrafast Peptide Label-Free Quantification with FlashLFQ. J Proteome Res2018, 17 (1), 386–391. [PubMed: 29083185]
6. Lu L; Millikin RJ; Solntsev SK; Rolfs Z; Scalf M; Shortreed MR; Smith LM, Identification of MS-Cleavable and Noncleavable Chemically Cross-Linked Peptides with MetaMorpheus. J Proteome Res2018, 17 (7), 2370–2376. [PubMed: 29793340]
7. Lu L; Riley NM; Shortreed MR; Bertozzi CR; Smith LM, O-Pair Search with MetaMorpheus for O-glycopeptide characterization. Nat Methods2020, 17 (11), 1133–1138. [PubMed: 33106676]
8. Rolfs Z; Millikin RJ; Smith LM, An Algorithm to Improve the Speed of Semi- and Non-Specific Enzyme Searches in Proteomics. Current Bioinformatics2020.
9. Miller RM; Millikin RJ; Hoffmann CV; Solntsev SK; Sheynkman GM; Shortreed MR; Smith LM, Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data. J Proteome Res2019, 18 (9), 3429–3438. [PubMed: 31378069]
10. Storey JD; Tibshirani R, Statistical significance for genomewide studies. Proc Natl Acad Sci U S A2003, 100 (16), 9440–5. [PubMed: 12883005]
11. Kall L; Storey JD; MacCoss MJ; Noble WS, Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res2008, 7 (1), 40–4. [PubMed: 18052118]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

12. Keller A; Nesvizhskii AI; Kolker E; Aebersold R, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem2002, 74 (20), 5383–92. [PubMed: 12403597]
13. Anderson DC; Li W; Payan DG; Noble WS, 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 Res2003, 2 (2), 137–46. [PubMed: 12716127]
14. Kall L; Canterbury JD; Weston J; Noble WS; MacCoss MJ, Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods2007, 4 (11), 923–5. [PubMed: 17952086]
15. Kall L; Storey JD; MacCoss MJ; Noble WS, Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res2008, 7 (1), 29–34. [PubMed: 18067246]
16. Kall L; Storey JD; Noble WS, Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics2008, 24 (16), i42–8. [PubMed: 18689838]
17. The M; MacCoss MJ; Noble WS; Kall L, Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom2016, 27 (11), 1719–1727. [PubMed: 27572102]
18. Halloran JT; Zhang H; Kara K; Renggli C; The M; Zhang C; Rocke DM; Kall L; Noble WS, Speeding Up Percolator. J Proteome Res2019, 18 (9), 3353–3359. [PubMed: 31407580]
19. Holzinger A, Data Mining with Decision Trees: Theory and Applications. Online Inform Rev2015, 39 (3), 437–438.
20. Choi H; Ghosh D; Nesvizhskii AI, Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res2008, 7 (1), 286–92. [PubMed: 18078310]
21. Granholm V; Noble WS; Kall L, A cross-validation scheme for machine learning algorithms in shotgun proteomics. BMC Bioinformatics2012, 13Suppl 16, S3.
22. Granholm V; Noble WS; Kall L, On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. J Proteome Res2011, 10 (5), 2671–8. [PubMed: 21391616]
23. Bassani-Sternberg M; Braunlein E; Klar R; Engleitner T; Sinitcyn P; Audehm S; Straub M; Weber J; Slotta-Huspenina J; Specht K; Martignoni ME; Werner A; Hein R; D HB; Peschel C; Rad R; Cox J; Mann M; Krackhardt AM, Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun2016, 7, 13404. [PubMed: 27869121]
24. Schaffer LV; Rensvold JW; Shortreed MR; Cesnik AJ; Jochem A; Scalf M; Frey BL; Pagliarini DJ; Smith LM, Identification and Quantification of Murine Mitochondrial Proteoforms Using an Integrated Top-Down and Intact-Mass Strategy. J Proteome Res2018, 17 (10), 3526–3536. [PubMed: 30180576]
25. Cesnik AJ; Miller RM; Ibrahim K; Lu L; Millikin RJ; Shortreed MR; Frey BL; Smith LM, Spritz: A Proteogenomic Database Engine. J Proteome Res2020.
26. Sheynkman GM; Shortreed MR; Frey BL; Smith LM, Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol Cell Proteomics2013, 12 (8), 2341–53. [PubMed: 23629695]
27. Cesnik AJ; Miller RM; Ibrahim K; Lu L; Millikin RJ; Shortreed MR; Frey BL; Smith LM, Spritz: A Proteogenomic Database Engine. bioRxiv2020.
28. Sherry ST; Ward M; Sirotkin K, dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res1999, 9 (8), 677–9. [PubMed: 10447503]
29. Jiang H; Lei R; Ding SW; Zhu S, Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics2014, 15, 182. [PubMed: 24925680]
30. Kim D; Paggi JM; Park C; Bennett C; Salzberg SL, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol2019, 37 (8), 907–915. [PubMed: 31375807]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

31. DePristo MA; Banks E; Poplin R; Garimella KV; Maguire JR; Hartl C; Philippakis AA; del Angel G; Rivas MA; Hanna M; McKenna A; Fennell TJ; Kernytsky AM; Sivachenko AY; Cibulskis K; Gabriel SB; Altshuler D; Daly MJ, A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet2011, 43 (5), 491–8. [PubMed: 21478889]

32. McKenna A; Hanna M; Banks E; Sivachenko A; Cibulskis K; Kernytsky A; Garimella K; Altshuler D; Gabriel S; Daly M; DePristo MA, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res2010, 20 (9), 1297–303. [PubMed: 20644199]

33. Cingolani P; Platts A; Wang le L; Coon M; Nguyen T; Wang L; Land SJ; Lu X; Ruden DM, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin)2012, 6 (2), 80–92. [PubMed: 22728672]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

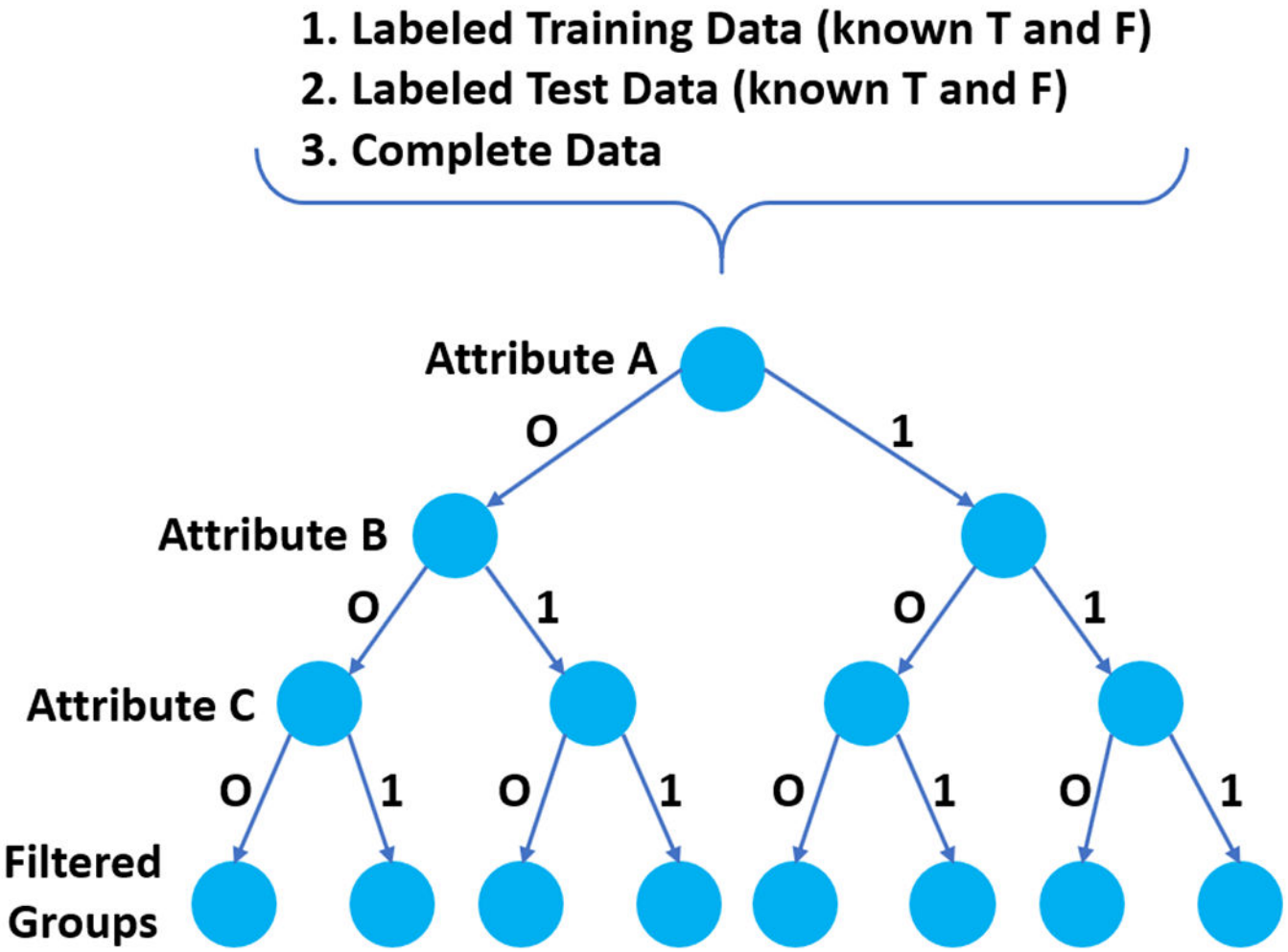


Figure 1.

Binary decision trees are created to classify subjects according to various attributes into one of two groups (e.g. true 'T' or false 'F'). There are three stages in the process: creation of the BDT, using training data; testing of the BDT performance, using test data; and finally, application of the BDT to the complete set of data. There is no overlap in membership between the training and test sets. The simplified example of a BDT shown in this figure uses three different attributes to classify all the incoming subjects into eight groups. The three attributes provide a total of seven different gates that effectively shuttle subjects into different groups (leaves). Each subject (e.g. a PSM) is evaluated with respect to each attribute. In this figure, classification is shown as a 1 or 0, which works for yes or no attributes (e.g. Is this a variant peptide sequence?). However, this is an oversimplification for attributes that are continuous variables (e.g. intensity), and in these cases the classification involves regression. The order of the attributes is chosen to maximize separation between the two groups at each stage using training data. The fraction of false subjects in the leaves of each branch of the tree provides the probability of false (posterior error probability) for all subjects in that leaf once the binary decision tree has been applied to the data. Known false test subjects are PSMs marked as decoys. Attribute 'A' is selected automatically to maximize separation between true and false. Attribute B is chosen next, using the same

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

guiding principle, which is maximization of the separation between true and false. The second attribute can be different between different branches of the tree. Each attribute can only be used once along a branch. The branch may terminate if 100% purity is achieved at any level. Once the tree is constructed, it is evaluated using a similarly sized, separate set of labeled test data. At this point, the BDT trained and ready to assign group membership to each data point and to compute the posterior error probability.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

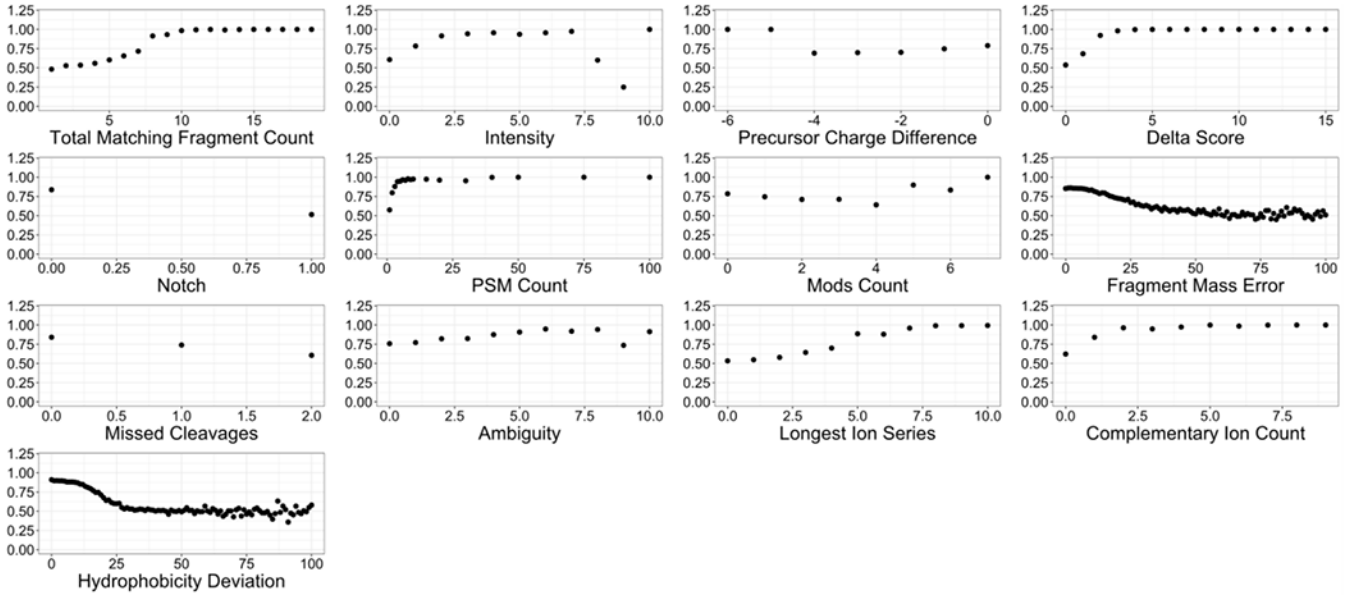


Figure 2.

The vertical axis reports the fraction of true positives across the range of observed values for the attributes used in the BDT. The units and ranges for the x-axes are arbitrarily chosen to allow the full range of fractions to be shown and should not be interpreted (see Supplement for explanation of the axes of each attribute). Note: a graph for the peptide variant feature is not shown.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

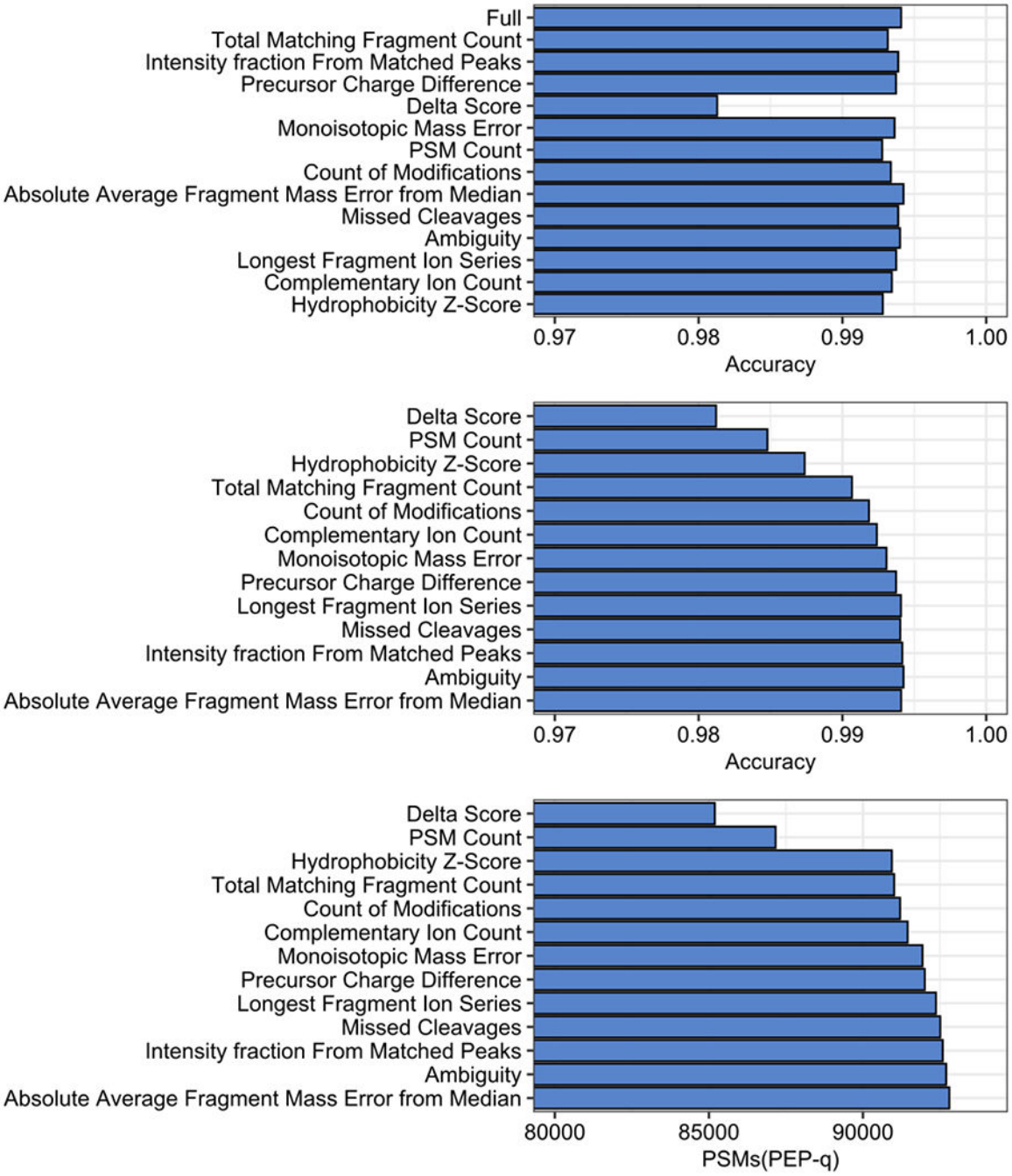


Figure 3.

(top) The accuracy of BDT models trained when missing the labelled attribute. When compared to the Full accuracy, this demonstrates the loss in value when not including the labeled attribute. (middle) The accuracy results of BDTs constructed one attribute at a time. Delta Score had the largest impact on accuracy as shown in top bar chart. Therefore, it was the first feature added to generate the data for middle bar chart. PSM count had the 2nd largest impact on accuracy in the top bar chart. Therefore, it was the 2nd feature added to generate the data for middle bar chart. This was repeated for all attributes.Each additional

added attribute improves accuracy. (bottom) The number of PSMs found at a PEP q-value < 0.01 increases with each added attribute. Here the BDT training began only with the Delta Score attribute. Then one-by-one, moving top to bottom, we added the labeled attribute and recompleted the search.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Definitions for attributes used in the binary decision tree.

Attribute	Definition
Absolute Average Fragment Mass Error from Median	Difference between the average fragment error (ppm) for a given PSM and the average fragment error for all PSMs
Ambiguity	Count of PSMs matching a single spectrum with identical MetaMorpheus score (+/− 1E-09)
Complementary Ion Count	Count of complementary fragment ion pairs where N- and C-terminal peptide fragments from the same backbone cleavage are observed.
Delta Score	Difference in MetaMorpheus score between the current PSM and the next best scoring PSM
Hydrophobicity Z-Score	The number of standard deviations the computed hydrophobicity/mobility a PSM differs compared with other PSMs eluting within two minutes.
Fraction of Spectrum Intensity from Matched Peaks	Normalized fraction of spectrum intensity assigned to the matched fragment ions of the PSM
Peptide Contains Amino Acid Variant	If the matched peptide contains a designated amino acid variant
Longest Fragment Ion Series	Count of consecutive peptide backbone cleavages annotated by either an N- or C-terminal fragment
Missed Cleavages Count	Count of missed proteolytic cleavage events for the peptide matched in the PSM
Count of Modifications	Count of peptide posttranslational chemical modifications
Monoisotopic Mass Error	Degree of missed monoisotopic error between experimental parent mass and computed theoretical mass (deconvolution error)
Precursor Charge Difference to Precursor Charge Mode	Integer difference between the charge state of the observed PSM and the mode for all PSMs
Peptide Spectral Match Count	Count of PSMs for the save full peptide sequence including any modifications
Total Matching Fragment Count	Count of all matched fragment ions

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Metrics of model performance. Portions of the text included below were adapted from https://github.com/dotnet/docs/blob/master/docs/machine-learning/resources/metrics.md, the original source of the BDT algorithm used in MetaMorpheus BDT.

Metric	Definition	What to look for
Accuracy	The proportion of correct predictions with a test data set. It is the ratio of number of correct predictions to the total number of input samples.	The closer to 1.00, the better.
Area Under Curve	Measures the area under the curve created by sweeping the true positive rate vs. the false positive rate.	The closer to 1.00, the better.
Area Under Precision Recall Curve	Area under the curve of a Precision-Recall curve, a measure of success of prediction when the classes are imbalanced.	The closer to 1.00, the better.
F1 Score	F1 score is the harmonic mean of the precision and recall.	The closer to 1.00, the better.
Log Loss	Logarithmic loss measures the performance of a classification model where the prediction input is a probability value between 0.00 and 1.00.	The closer to 0.00, the better.
Log Loss Reduction	The advantage of the classifier over a random prediction.	Ranges from -inf and 1.00, where 1.00 is perfect predictions and 0.00 indicates mean predictions.
Positive Precision	The proportion of correctly predicted positive instances among all the positive predictions.	The closer to 1.00, the better.
Positive Recall	The proportion of correctly predicted positive instances among all the positive instances.	The closer to 0.00, the better.
Negative Precision	The proportion of correctly predicted negative instances among all the negative predictions.	The closer to 0.00, the better.
Negative Recall	The proportion of correctly predicted negative instances among all the negative instances.	The closer to 0.00, the better.
Count of Ambiguous Peptides Removed	Peptide assignments with the same MetaMorpheus score resolved through application of the BDT.	Higher numbers are better.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3.

Figures of merit for search vignettes

	Bottom-Up	Top-Down	Non-Specific (HLA)	Proteogenomic
Accuracy	0.9941	0.9959	0.9993	0.9919
Area Under the Curve	0.9995	0.9997	0.9995	0.9967
Area Under Precision Recall Curve	0.9993	0.9999	0.9990	0.9973
F1 Score	0.9936	0.9970	0.9973	0.9911
Log Loss	0.0297	0.0305	0.0077	0.0475
Log Loss Reduction	0.9702	0.9664	0.9861	0.9523
Positive Precision	0.9942	0.9963	0.9970	0.9926
Positive Recall	0.9931	0.9976	0.9977	0.9897
Negative Precision	0.9939	0.9949	0.9997	0.9913
Negative Recall	0.9950	0.9923	0.9996	0.9938
Count of Ambiguous Peptides Removed	210	506	115	786