# Leveraging Bottom-Up Search Results to Build Sample Specific Top-Down Databases

Mick Greer[1], Peter Verhaert[2], Kenneth Verheggen[2], Ken Durbin[3], Joe Greer[3], Ryan Fellers[3], Rich Leduc[2]    1 ThermoFisher Scientific, Austin, TX; 2 ProteoFormiX, Beerse, BE; 3 Proteinaceous, Evanston, IL.

## ABSTRACT

**Purpose:** Demonstrate the value of a combined bottom-up (BU) and top-down (TD) LC-MS strategy.

**Material/Methods:** Focus was on proteoform identification in samples of an under-annotated proteome (the archaeon *Sulfolobus islandicus*), and a sample from the well annotated *Homo sapiens* proteome was analyzed using the same workflow for comparison. Data were processed with Thermo Scientific™ Proteome Discoverer™ software using Thermo Scientific™ ProSightPD™ nodes.

**Results:** Adding modifications discovered via bottom-up (tryptic) peptide analysis to a top-down database significantly improves proteoform identifications from top-down analyses.

## INTRODUCTION

Bottom-up proteomics effectively infers the presence of thousands of proteins in a single experiment. However, peptide level measurements often fail to identify the relevant proteoforms in a given sample. When protein modifications are crucial for biological activity, bottom-up analyses remain inadequate to confidently dissect a targeted biological process to its molecular level.

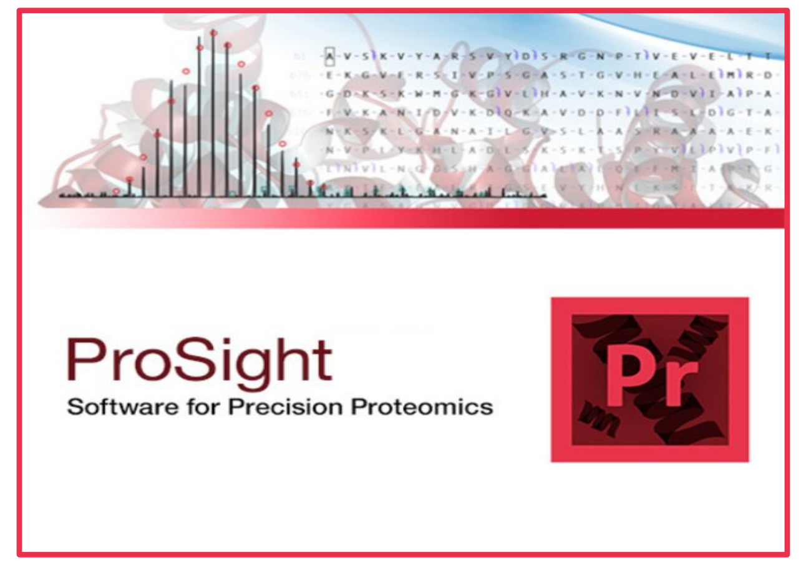
Top-down proteomics identifies intact proteoforms but, compared to (tryptic) peptide identifications, tandem MS analyses of large proteoform ions have poor efficiency. Successful top-down identifications, therefore, benefit from well annotated databases, which include knowledge of PTMs, mutations, and other modifications. Such databases do not yet exist for most species. Previously several other reports have shown the benefits of utilizing peptide information to enrich proteoform results[1,2]. We demonstrate a combined bioinformatics strategy using bottom-up data to compile sample specific annotated databases for subsequent top-down analysis within the same software platform.

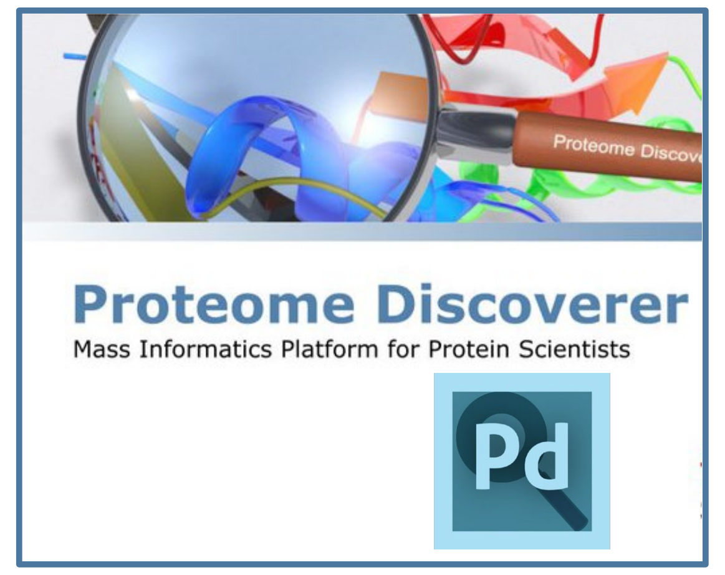
## MATERIALS AND METHODS

### Sample Preparation

To evaluate our strategy, two complementary bottom-up (BU) and top-down (TD) data sets were used. One human protein sample was selected to represent a highly studied and well-annotated eukaryote, the other one represents a much less-studied prokaryote species.

[1] A purposely in-house collected BU/TD tear protein dataset from *Homo sapiens*. BU and TD analyses were done in replicate on the very same tear fluid samples from a Schirmer strip-based collection method described before[3]. LC-MSMS analyses were performed by on-line EASY-nLC™ (C18 and C5 chromatography for BU and TD analyses respectively) coupled to a Thermo Scientific™ QExactive™ Plus Hybrid Quadrupole-Orbitrap™ mass spectrometer. [2] A 2nd sample from *Sulfolobus islandicus* (a thermoacidophylic archaeon with a nearly unannotated proteome). BU and TD *S. islandicus* samples were pre-fractionated before separation and analysis on a Thermo Scientific™ UltiMate™ 3000 nano-LC system and a Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer.





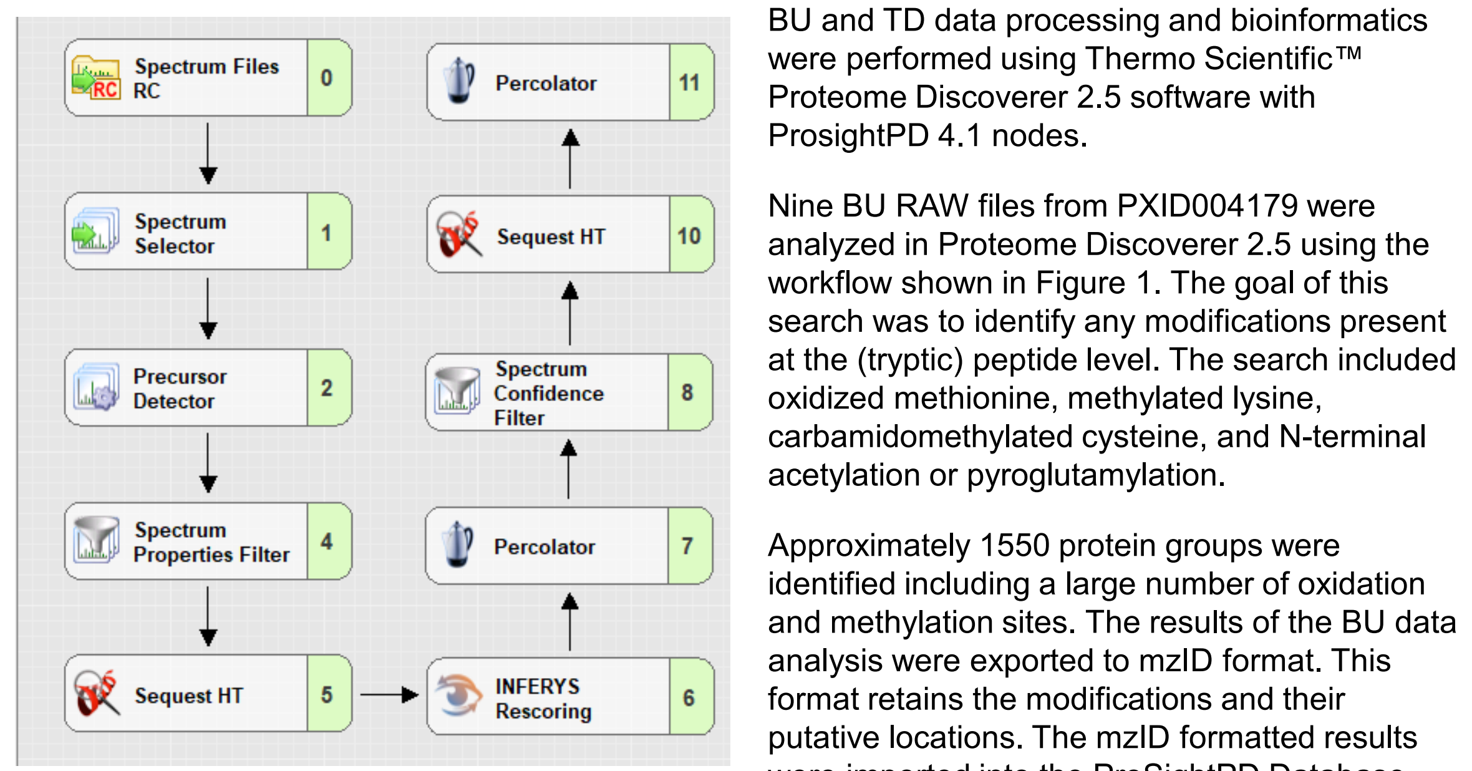


Figure 1. Bottom up data analysis workflow in Proteome Discoverer 2.5

### Data Analysis

BU and TD data processing and bioinformatics were performed using Thermo Scientific™ Proteome Discoverer 2.5 software with ProsightPD 4.1 nodes.

Nine BU RAW files from PXID004179 were analyzed in Proteome Discoverer 2.5 using the workflow shown in Figure 1. The goal of this search was to identify any modifications present at the (tryptic) peptide level. The search included oxidized methionine, methylated lysine, carbamidomethylated cysteine, and N-terminal acetylation or pyroglutamylation.

Approximately 1550 protein groups were identified including a large number of oxidation and methylation sites. The results of the BU data analysis were exported to mzID format. This format retains the modifications and their putative locations. The mzID formatted results were imported into the ProSightPD Database Manager and converted to a top-down database.

In turn six TD data files from PXID003074 were analyzed using the ProSightPD 4.1 workflow shown in Figure 2. These data were analyzed using three different databases in the same workflow to evaluate the utility of making a top-down database using bottom results. The databases were created as follows:

[1] FASTA database only (sequences w/o any modifications);

[2] Annotated database from Proteome Discoverer BU peptide results (only modifications found in BU search);

[3] Database from [2] with Oxi M and Methyl K set to variable on all methionine and lysine residues.

Four TD data files collected from tear fluid were analyzed using the workflow in Figure 2. The resulting data were processed using three different databases similar as above. However, in this case UniProt, a well annotated *H. sapiens* proteome database, was used in .xml format instead of the unannotated .FASTA (no PTMs or mods) database employed for *S. islandicus.*

[1] .xml database from UniProt (containing sequences and any annotated modifications);

[2] Annotated database from Proteome Discoverer peptide results (only modifications found in BU search);

[3] Database from [2] with Oxi M, Deamidated N+Q, and PyroGlu set to variable on all M,N,Q residues respectively.

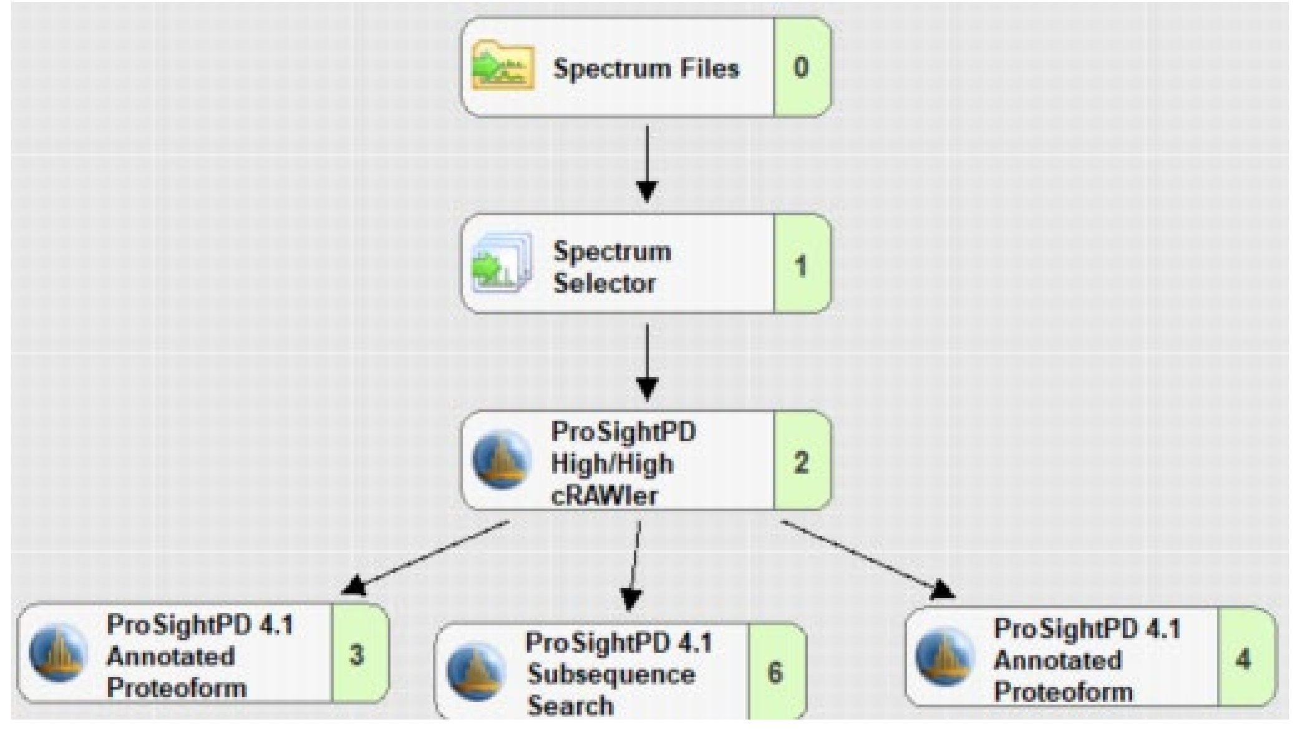

Spectrum Files
0
Spectrum Selector
1
ProSightPD High/High cRAWler
2
ProSightPD 4.1 Annotated Proteoform
3
ProSightPD 4.1 Subsequence Search
6
ProSightPD 4.1 Annotated Proteoform
4

Figure 2. Top-down analysis in Proteome Discoverer 2.5 using ProSightPD 4.1 nodes

## Results

### Increasing Proteoform Search Space to Reflect Sample Complexity

Figure 3 shows a representative isoform entry of M9U8W4 in Database Manager from database 1 (FASTA sequence only.) This database includes only ~2000 proteoforms from the canonical sequences. At the time of search these will be expanded to include N-terminal acetylation, and removal of methionine. Increasing the search space by addition of the BU annotations increased the search space to ~5E13 proteoforms, which yields an increase in the number of M9UW4 proteoforms from 1 to 8 (Figure 4.) Adding the variable methylation of all lysine residues and oxidation of all methionine residues in the database increased the search space to ~2E26 proteoforms, which results in an increased number of M9UW4 proteoforms from 60 (Figure 5.)

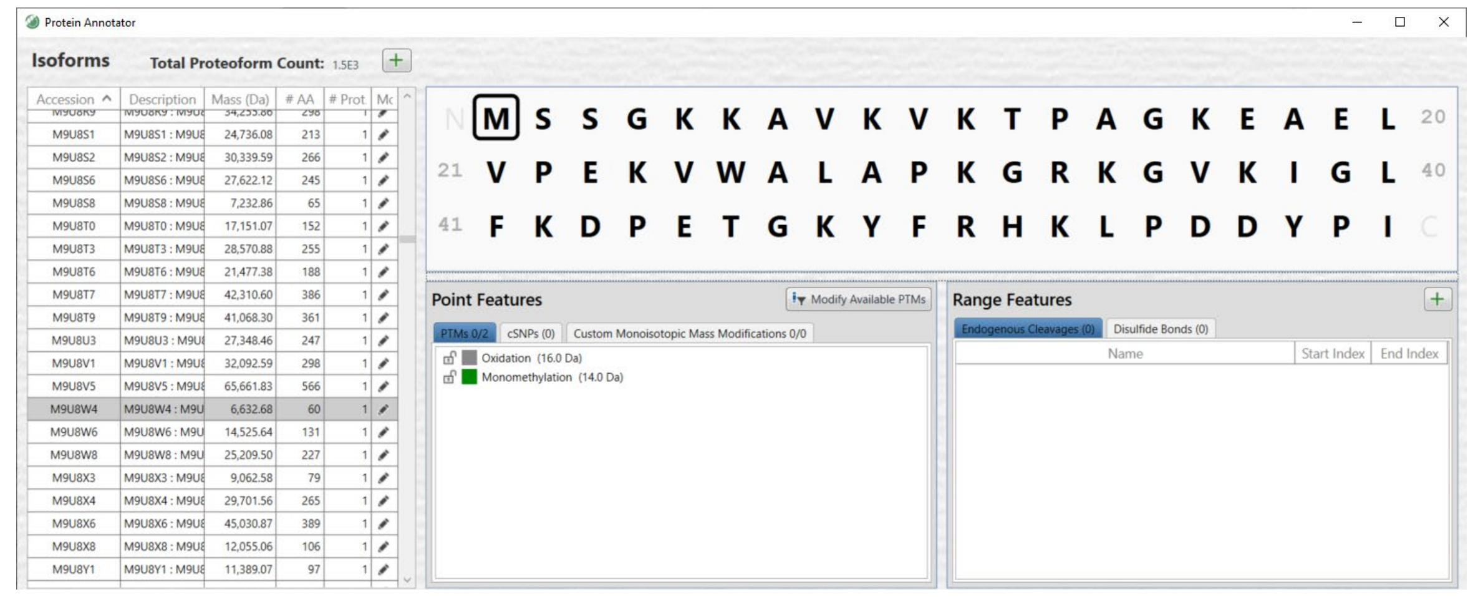

Protein Annotator
Isoforms
Total Proteoform Count: 1.5E3
Accession
Description
Mass (Da)
# AA
# Prot
M S S G K K A V K V K T P A G K E A E L 20
21 V P E K V W A L A P K G R K G V K I G L 40
41 F K D P E T G K Y F R H K L P D D Y P I
Point Features
Range Features

Figure 3. Representative isoform from the FASTA only top down database

Addition of PTM annotation from Bottom Up results increases proteoform search space to 5E13

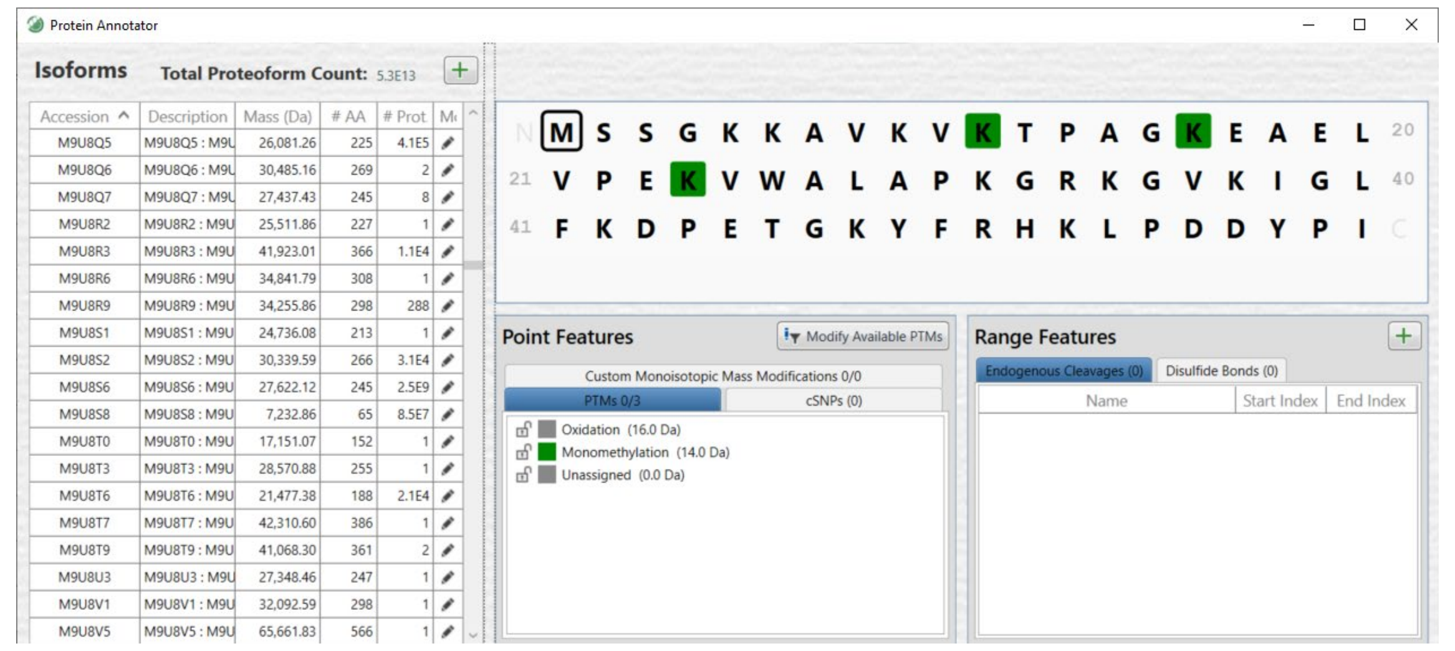

Protein Annotator
Isoforms
Total Proteoform Count: 5.3E13
Accession
Description
Mass (Da)
# AA
# Prot
M S S G K K A V K V K T P A G K E A E L 20
21 V P E K V W A L A P K G R K G V K I G L 40
41 F K D P E T G K Y F R H K L P D D Y P I
Point Features
Range Features
Oxidation (16.0 Da)
Monomethylation (14.0 Da)
Unassigned (0.0 Da)

Figure 4. Representative isoform from the BU annotated top down database

Addition of Methylation of Lysine and Oxidation of Methionine increases proteoform search space to 1.5E26

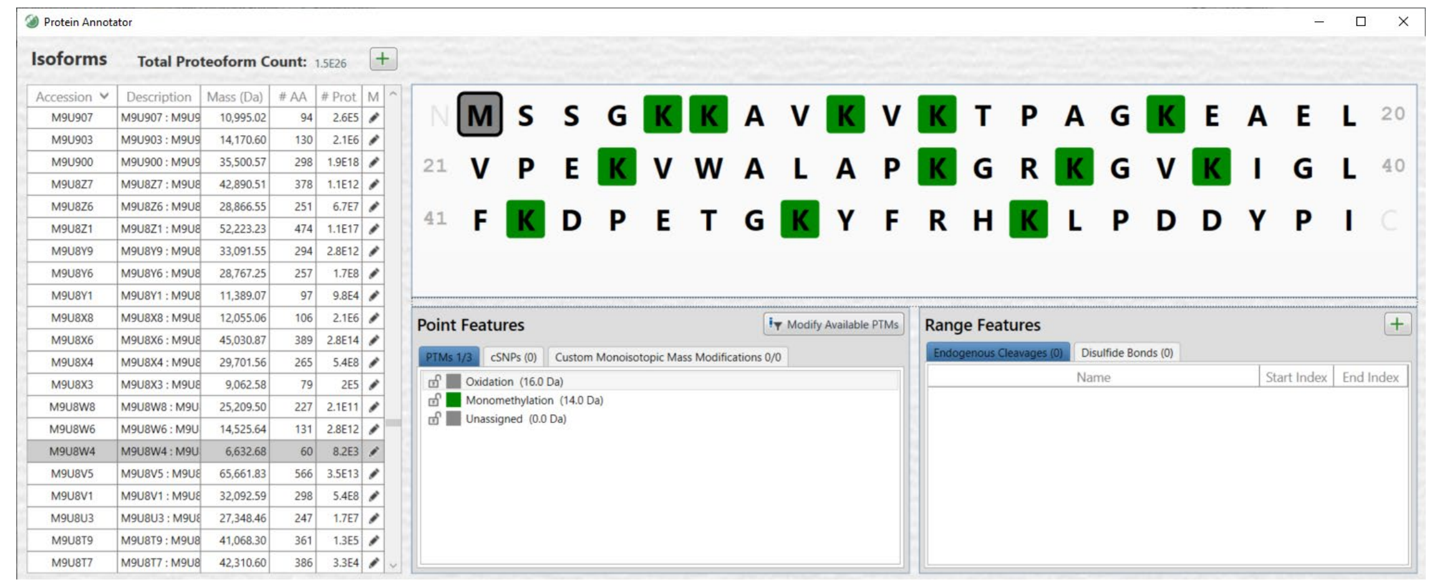

Protein Annotator
Isoforms
Total Proteoform Count: 1.5E26
Accession
Description
Mass (Da)
# AA
# Prot
M S S G K K A V K V K T P A G K E A E L 20
21 V P E K V W A L A P K G R K G V K I G L 40
41 F K D P E T G K Y F R H K L P D D Y P I
Point Features
Range Features

Figure 5. Representative isoform from the BU annotated + variable oxidation of M and methylation of K top down database

### Database Effect on Proteoform Identification Results

Figure 6 shows a representative result for M9U8W4 from a search using the FASTA only database. With this limited database only 4 proteoforms of M9U8W4 were identified. However, when the database containing BU annotated modifications was used, 15 proteoforms were identified (Figure7.) Finally, when variable methylation of lysine and oxidation of methionine were included over 100 proteoforms of M9U8W4 were identified (Figure 8).

These results demonstrate the importance of careful database creation for TD data analysis. Furthermore, this is an excellent example of a proteome with limited annotation in UniProt which can be analyzed at the peptide level and used to annotate a top-down database to significantly increase proteoform identifications. Ultimately however utilizing the BU informed database which included variable mods on lysine and methionine yielded the greatest number of identifications. This is likely due to the large proteoform search space. This also reflects the shortfall of peptide level analysis. Many lysine modification sites may be lost due to unfavorable tryptic cleavage (i.e. extremely short peptides) when lysine and arginine residues occur in close proximity. In this case adding potential methylations at all lysine sites can compensate.

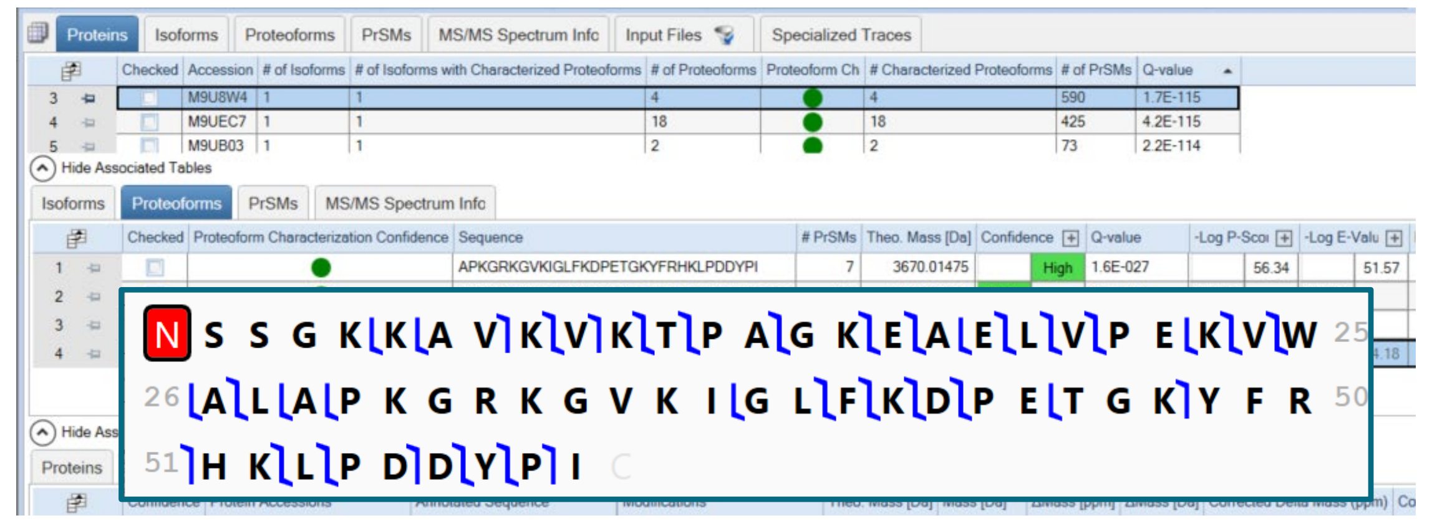


Figure 6. Representative search results using the FASTA only top down database

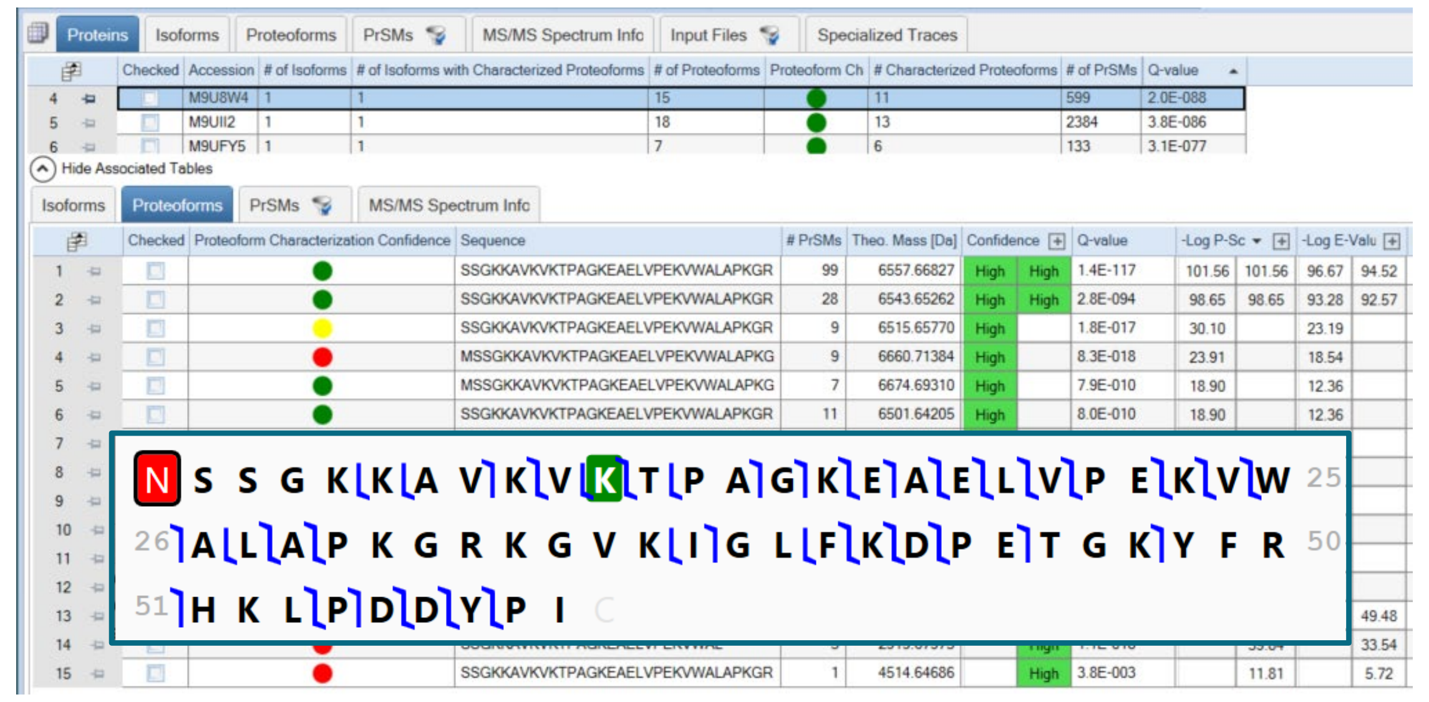


Figure 7. Representative search results using the BU annotated top down database

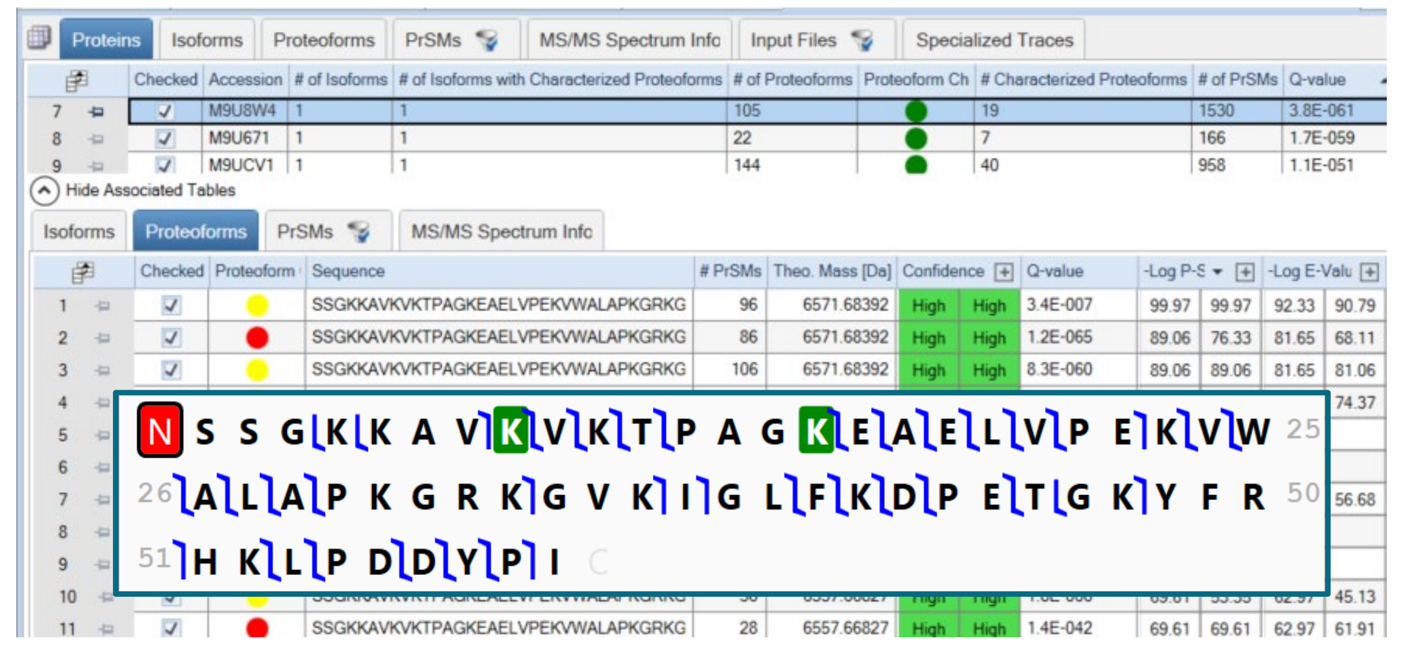
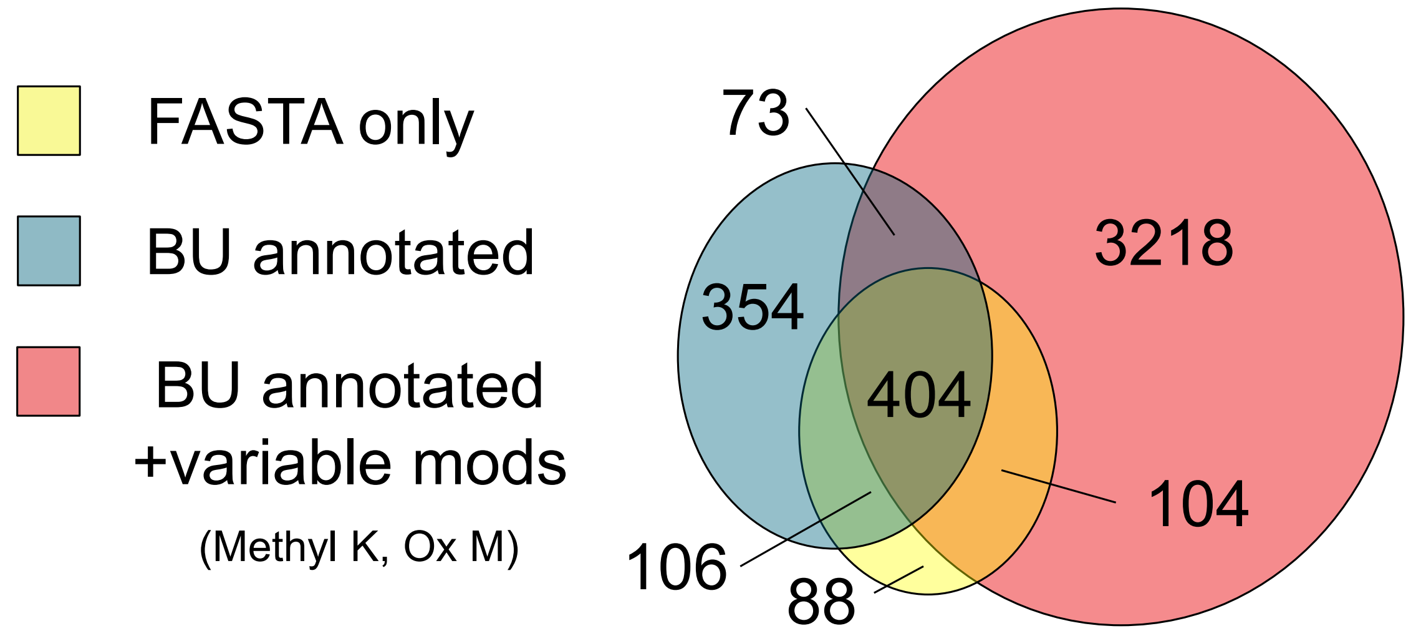

Proteins
Isoforms
Proteoforms
PrSMs
MS/MS Spectrum Info
Input Files
Specialized Traces
Hide Associated Tables
N S S G K K A V K V K T P A G K E A E L L V P E K V W 25
26 A L L A P K G R K G V K I G L F K D P E T G K Y F R 50
51 H K L L P D D Y P I
SSGKKAVKVKTPAGKEAELVPEKVWALAPKGRKG
High
1.4E-042

Figure 8. Representative search results using the BU annotated + variable oxidation of M and methylation of K top down database

Figure 6-8 (inset) shows a representative sequence coverage map from each result that illustrates the search space. As the search space increased, the number of identifiable modifications increases as well.

Figure 9 shows the total number proteoforms identified by searches using each top-down database and their overlap. These results were generated using a 1% FDR cutoff threshold at the proteoform level.



During the database search the proteoform search space is limited to a user defined number of PTMs to prevent excessive search times. In this case some proteoforms which were included in the BU annotated search were not included in the search containing variable mods (due to the increased number mods) resulting in the portion of unique proteoforms in the BU annotated group.

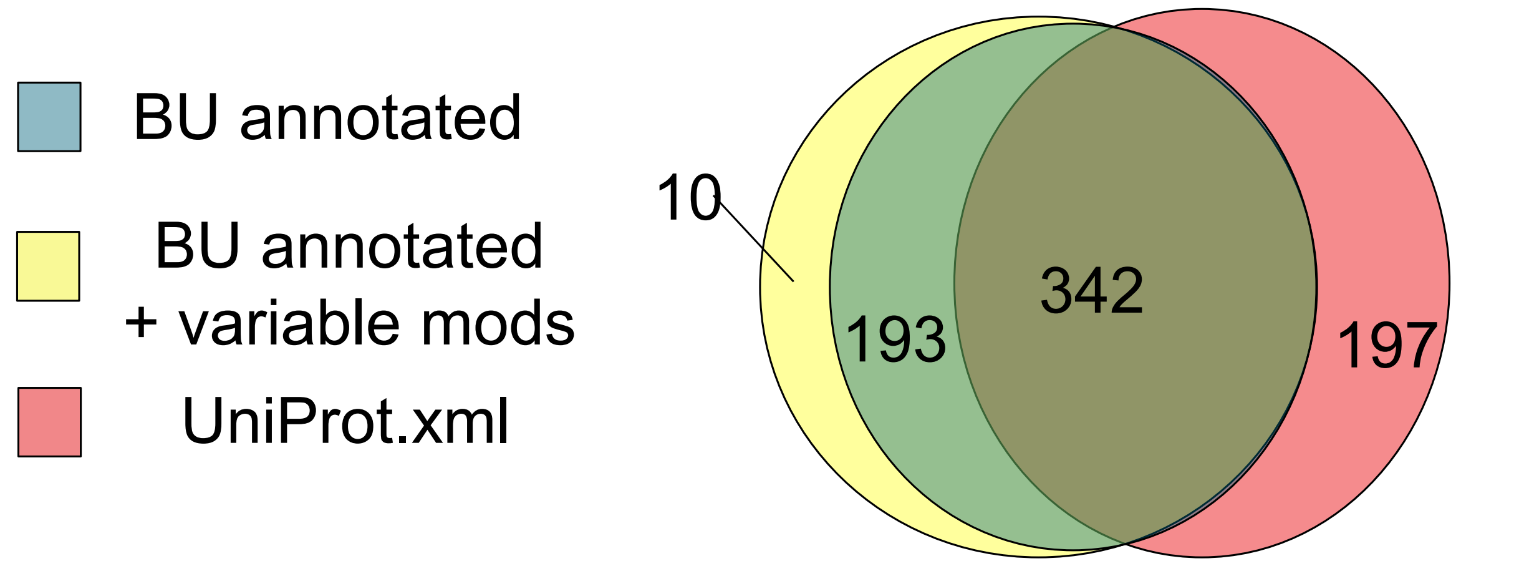


Figure 10 shows a comparison of TD results using different databases. Similar numbers of proteoforms were identified using the UniProt .xml and the BU annotated database with additional variable mods. From this we conclude that poorly studied organisms (Fig 9) benefit from additional peptide level data more so than well studied and annotated species such as human (Fig 10) where the proteoforms identified via this strategy are complementary with the annotations found in UniProt.

## CONCLUSIONS

- Database creation is a critical step in TD data analysis.
- TD database annotation via BU search results can be conveniently executed in ProSightPD and significantly improves proteoform identifications for poorly annotated proteomes.

## REFERENCES

1. Schaffer, L.V., Millikin, R.J., Shortreed, M.R., Scalf, M., & Smith, L.M. (2020) Improving Proteoform Identifications in Complex Systems Through Integration of Bottom-Up and Top-Down Data. *J Proteome Res*, *19:* 3510–3517.
2. Lima, D.B., Dupré, M., Duchateau, M., Gianetto, Q.G., Rey, M., Matondo, M., & Chamot-Rooke, J. (2021) ProteoCombiner: integrating bottom-up with top-down proteomics data for improved proteoform assessment. *Bioinformatics*, *37:* 2206–2208. [doi.org/10.1093/bioinformatics/btaa958]
3. Raus, P., Kumar-Raguraman, B., Pinkse, M., Verhaert, P. (2015) Bottom-up protein identifications from microliter quantities of individual human tear samples. Important steps towards clinical relevance. *EUPA Open Proteomics, 9: 8-13. [*doi/10.1016/j.euprot.2015.06.005]

## TRADEMARKS/LICENSING

PO66099 EN0921S