


# PMI Preview™

# User's Manual

September 2020

3.10

Protein Metrics Inc., Cupertino, California, USA

# CONTENTS


PMI Preview User's Manual ..... 1
Overview ..... 1
Introduction ..... 1
Main Screen ..... 1
Input Files ..... 2
Fixed Modifications ..... 2
Search options ..... 4
Save/Load Parameters ..... 5
Run Preview ..... 6
Preview Results ..... 7
Preview Summary Results ..... 9
Top Proteins ..... 9
Mass measurements ..... 9
Digestion ..... 10
Fixed modifications ..... 11
Most common variable modifications ..... 11
Detailed results ..... 11
Preview Detailed Results ..... 12
Table of Contents ..... 12
Representative Proteins ..... 12
Score distribution ..... 13
Mass measurement errors ..... 13
Cysteine ..... 14
Nonspecific cleavage ..... 14
Oxidation ..... 14
Chemical modifications ..... 14
Posttranslational modifications ..... 15
Computational information ..... 15
Main Menu Bar ..... 16
File Menu ..... 16
Edit Menu ..... 16
Help Menu ..... 17
Preview FAQs ..... 17

# PMI Preview User's Manual

## Overview

Protein Metrics Preview™ quickly samples MS data to measure mass errors, digestion specificity, and modifications in preparation for more thorough searches. Preview generates optimized parameters files for use in Protein Metrics Byonic™ searches. These parameters include precursor and fragment mass tolerances, digestion specificity, and allowed types of modifications. Preview runs fast initial searches to recalibrate precursor and fragment masses based on confident identifications. Preview samples the data set, but does not perform an exhaustive search, so it is not a substitute for a full database search program such as Byonic.

## Introduction

Preview was developed as a very fast searching program to guide parameter setting for subsequent full searches. Preview has only two required inputs: a set of MS/MS spectra in a variety of formats, and a protein database in FASTA format. The program computes appropriate settings for precursor and fragment tolerances, estimates the amount and type of nonspecific digestion, measures the prevalence of recognized modifications, and reports unrecognized/unusual mass shifts. Preview operates in a fraction of the time of a standard search program; for example, a search of 10,000 spectra against a database containing 100,000 protein sequences takes about 1 minute on a common desktop PC.

To achieve this speed, Preview uses several simplifying assumptions. The foremost assumption is that the 100 most detectable proteins is sufficient to represent the entire sample for the full menu of search parameters. Also, with some exceptions, Preview searches for unrelated modification types one at a time, thereby avoiding the combinatorial explosion and performance degradation of multiple modification searches. Finally, Preview's peptide identification algorithm takes shortcuts: it represents both predicted and observed peaks by integer masses, as a faster algorithm for candidate scoring.

Preview optionally recalibrates m/z measurements and outputs a new spectrum file (in *.mgf format) that can be used as input into Byonic. The user can choose to recalibrate precursor measurements, fragment measurements, or both.

Preview can benefit a proteomics laboratory in several ways. First, it can help the user choose parameter settings for subsequent full search by Byonic or any other search engine. Second, even if the user already knows the best parameter settings, m/z recalibration can improve sensitivity and specificity by increasing the scores of true matches relative to those of false matches. Finally, timely feedback on sample preparation artifacts and m/z errors can help improve laboratory practices and experimental reproducibility.

## Main Screen

Preview contains several sections:

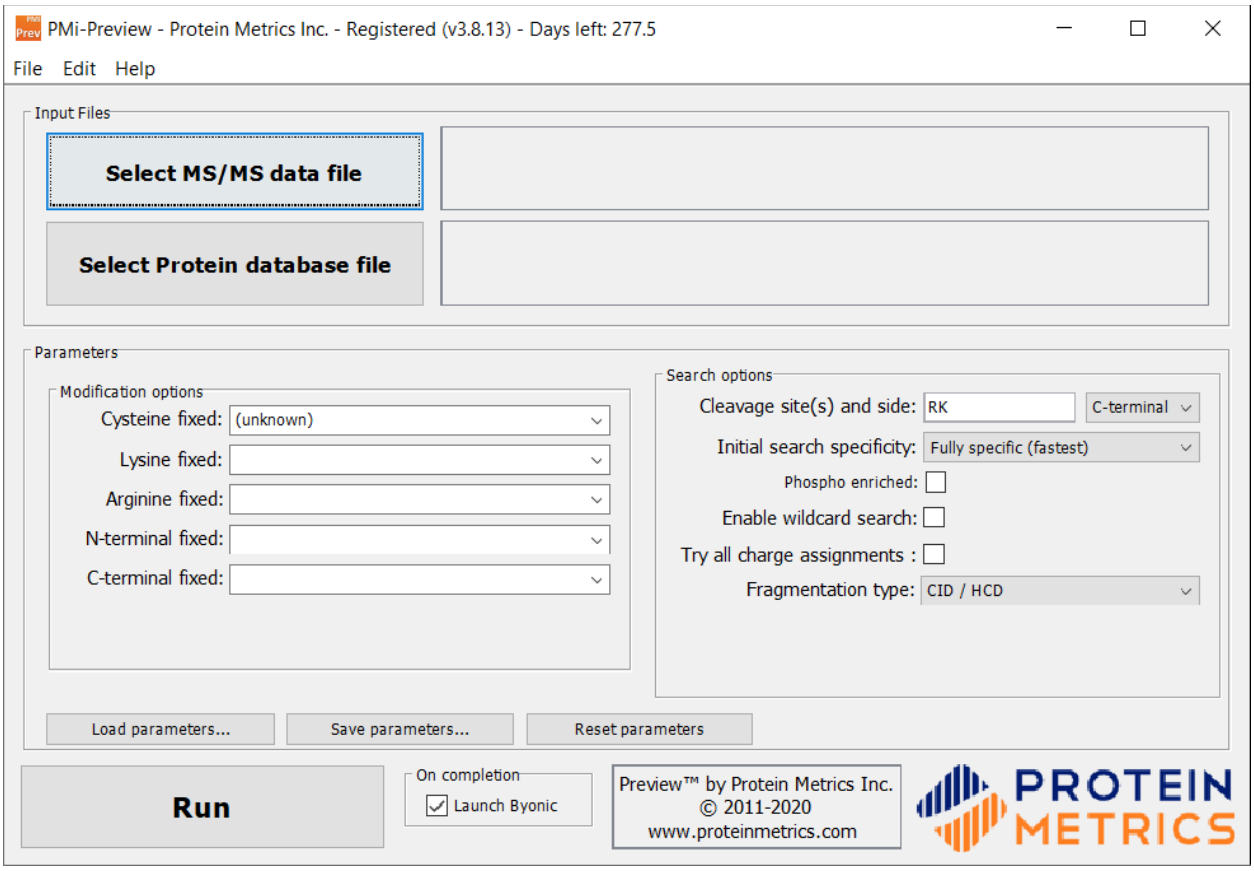



Figure 1: Initial Preview screen

## Input Files

The first section is **Input Files**. To load a sample file, click **Select MS/MS data file**, navigate to and select the MS file and click **Open**. Most of the sample data file types supported by PMI applications are supported by Preview. These include: Bruker: *.d, Thermo: *.raw, Sciex: *.wiff and *.wiff2, Agilent: *.d, Shimadzu *.lcd, as well as *.mgf, *.mzML and *.mzXML. Preview also supports PMI *.byspec2 files.

To load a protein database (FASTA) file, click **Select Protein database file**, navigate to and select the *.fasta or *.fa file and click **Open**. The two filenames are displayed in the cell at right:

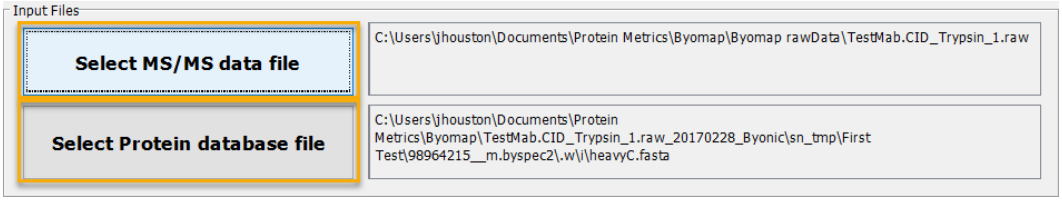



Figure 2: Input Files

An example fasta file is common.contaminants.fasta found in the directory: `C:\Program Files\ProteinMetrics\PMI-Suite\Tools\Preview\data`.

## Fixed Modifications

The **Parameters** section contains the **Modification options** section on the left:

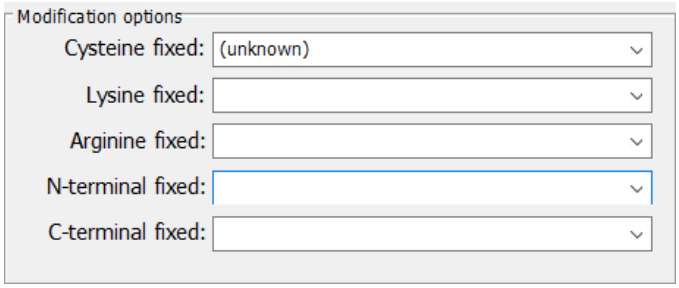



Figure 3: Modification options

One of the most difficult choices for parameter settings are which peptide modifications to allow. Some in vitro modifications are ubiquitous, occurring to some extent in almost all shotgun proteomics samples. Other modifications depend upon both the sample and its preparation and can vary unpredictably. Posttranslational modifications (PTMs) also vary from sample to sample and from protein to protein within a sample.

The supported fixed modifications and their available options are:

- **Cysteine fixed**

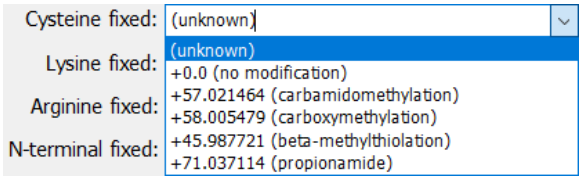



Figure 4: Cysteine fixed options

For standard cysteine treatments (+0, +46, +57, +58, and +71), this input is not always necessary. If the user selects “(unknown)”, Preview usually determines the correct cysteine treatment from the data.

- **Lysine fixed**

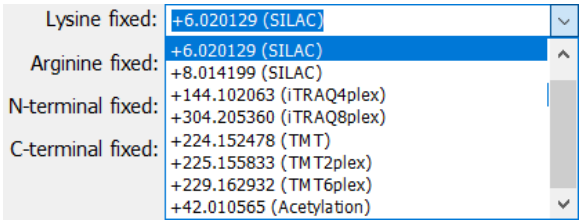



Figure 5: Lysine fixed options

- **Arginine fixed**

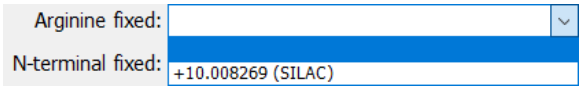



Figure 6: Arginine fixed options

- **N-terminal fixed**

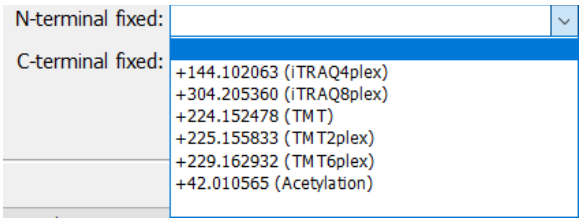



Figure 7: N-terminal fixed options

- **C-terminal fixed**

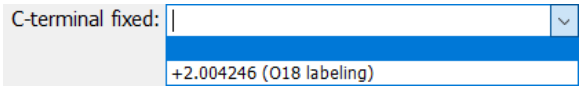



Figure 8: C-terminal fixed options

To enter a custom modification, enter the mass delta (starting with + or -) in daltons into the appropriate fixed modification cell.

To remove a modification selection, select the blank. In this default mode, Preview will assume 0 Da for Lysine, Arginine, N-terminal, and C-terminal, and will not test for other possibilities for these modifications.

Preview is good at finding modifications that are consistently found sample-wide (for example, in vitro modifications from sample handling and processing). However, searches may miss posttranslational modifications that are found only on a few proteins. Note that Preview does not search for glycopeptides. To enable searches of specific posttranslational modifications or glycopeptides, carefully review and adjust the Byonic parameters suggested by the Preview output.

## Search options

The **Search options** section, also in the Parameters section contain fragmentation, digestion and other search parameters:

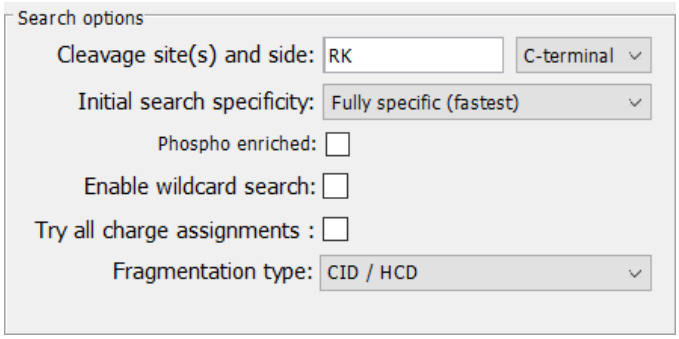



Figure 9: Search options

The Search options include:

- **Cleavage site(s) and side** set the point of cleavage and whether to cleave on the **C-terminal** or **N-terminal** side. Enter one-letter abbreviations for residues on either side of the cleavage point. (in this example, trypsin, on the C-terminal side of arginine or lysine). For a broader search, use nonspecific cleavage at one or both termini. Nonspecific digestion can vary from negligible to ubiquitous depending upon endogenous peptidases, and missed cleavages vary widely depending upon sample processing conditions.
- **Initial search specificity** sets the level of specificity, and thus the search speed. The options are:

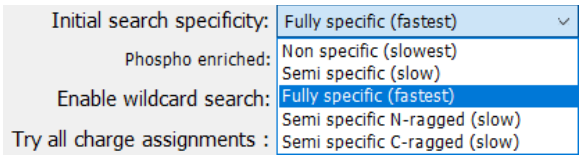



Figure 10: Search specificity options

Fully specific searches are recommended for all digested samples. Nonspecific initial searches may perform better for undigested (peptidomic) samples. In Fully specific searches (the default), both peptide termini must agree with the input digestion cleavages. In N-ragged searches, only the C-terminus must agree, and in C-ragged searches, only the N-terminus must agree. In Semi specific searches, one the two termini can disagree. In Non specific searches, both termini can disagree.

- Check **Phospho enriched** to optimize Preview when the sample is composed predominantly of phosphopeptides.
- Check **Enable wildcard search** to enable searches of the spectra with a wild-card modification. When it is checked, the program performs a blind modification search that tries each integer mass shift from -50 to +150 on any one residue. The mass of the modification will be reported to the accuracy of the precursor mass.
- When **Try all charge assignments** is checked, the charge assignments in the spectrum file are ignored. The program runs every spectrum using z = +1, +2, +3 for each CID spectrum, and z = +2, +3, +4 for each ETD spectrum. You can run Preview twice, once with this box checked and once without, to test the reliability of the charge assignments.
- The **Fragmentation type** options are:

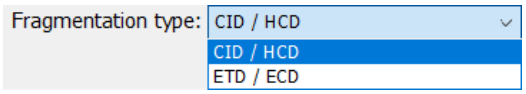



Figure 11: Fragmentation type options

CID / HCD represent b and y ions, and ETD / ECD represent c and z ions.

## Save/Load Parameters

At the bottom of the Parameters section are three buttons to manage the set of parameters:

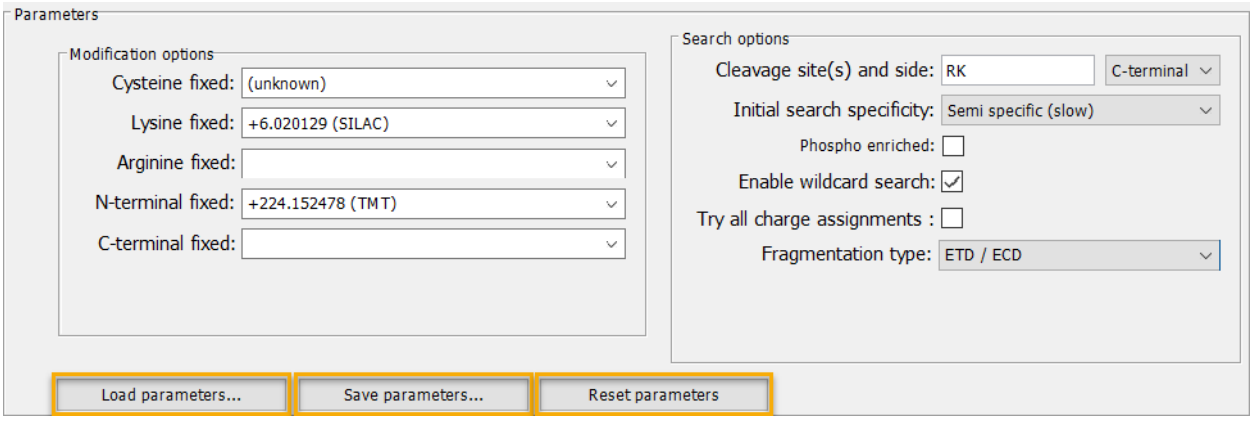



Figure 12: Parameter management buttons

- **Load parameters** open a collection of parameters (*.prv file) saved in a prior Preview session.

- **Save parameters** saves the current parameters as a Preview parameter (*.prv) file.
- **Reset parameters** sets the current parameters back to the default settings.

## Run Preview

Before running Preview, the user has the option to automatically open Byonic and load it with the parameters used in Preview. To set this option, check **On completion Launch Byonic**.

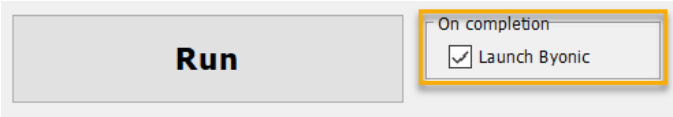



Figure 13: Run button and Launch Byonic checkbox

If an output folder has not been set before the Preview run, the user will be prompted to set one. To set an output folder, choose **Edit > Preferences**:, click **Select Folder**, navigate to and select the desired folder and click **Open**.

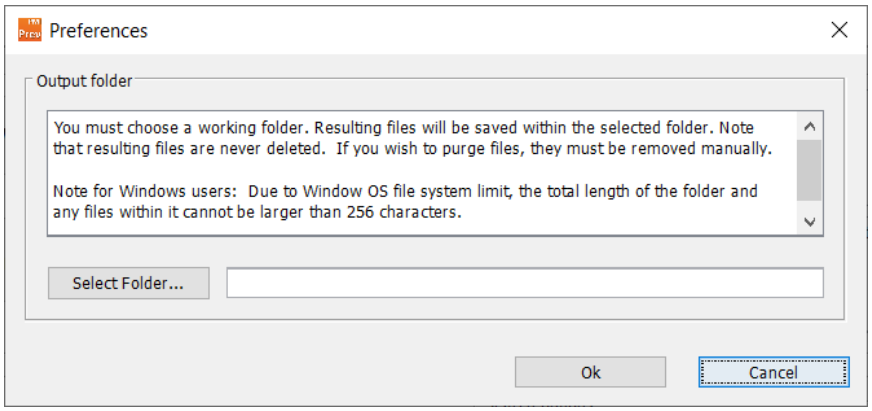



Figure 14: Setting output folder in Preferences

When all parameters and the output folder are set, click **Run** to start Preview.

A status window opens to show that the run is in progress:

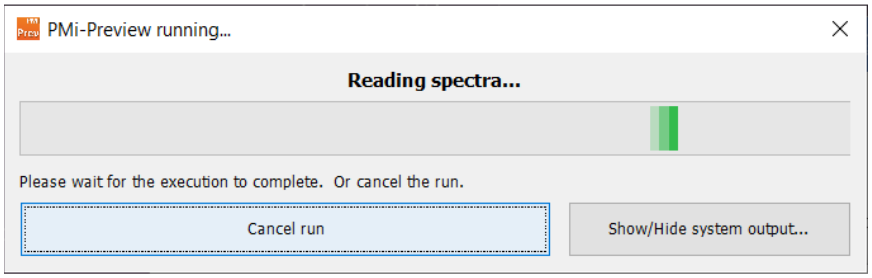



Figure 15: Preview run status

To cancel the process, click **Cancel run**. To view details of the process, click **Show/Hide system output**.

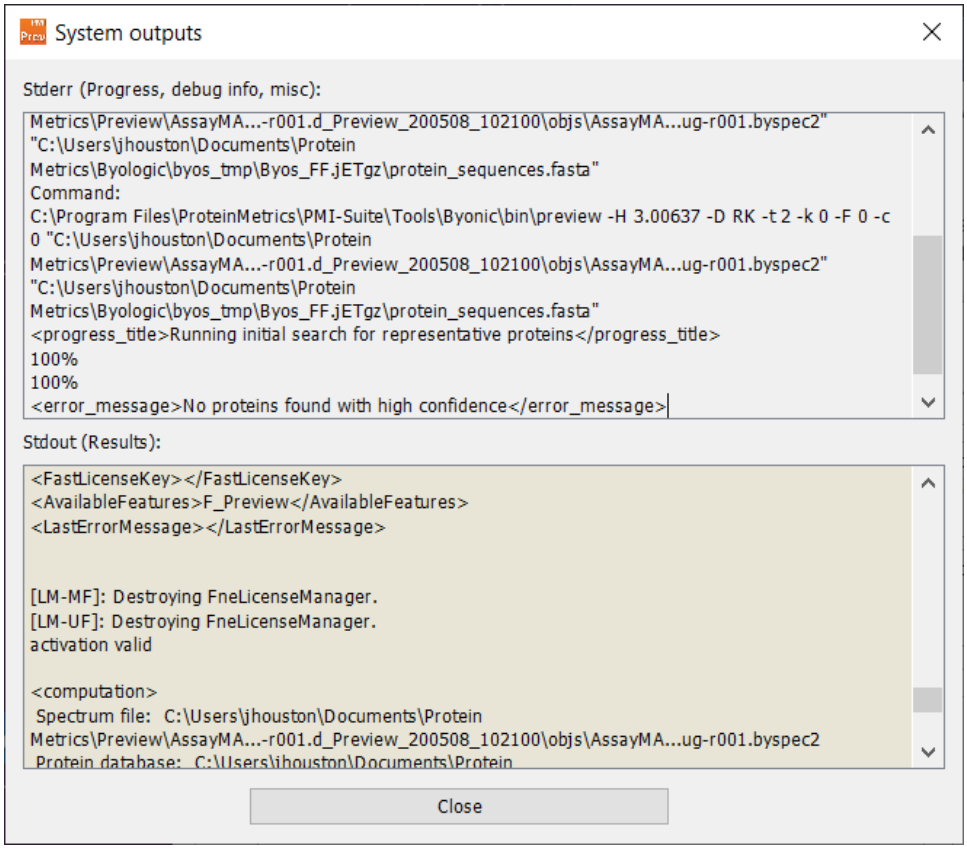



Figure 16: Process System outputs

## Preview Results

After a successful run, Preview displays the results as an html page in the default web browser. The initial page shows the summary results, also displayed when **Summary** is clicked in the title bar:

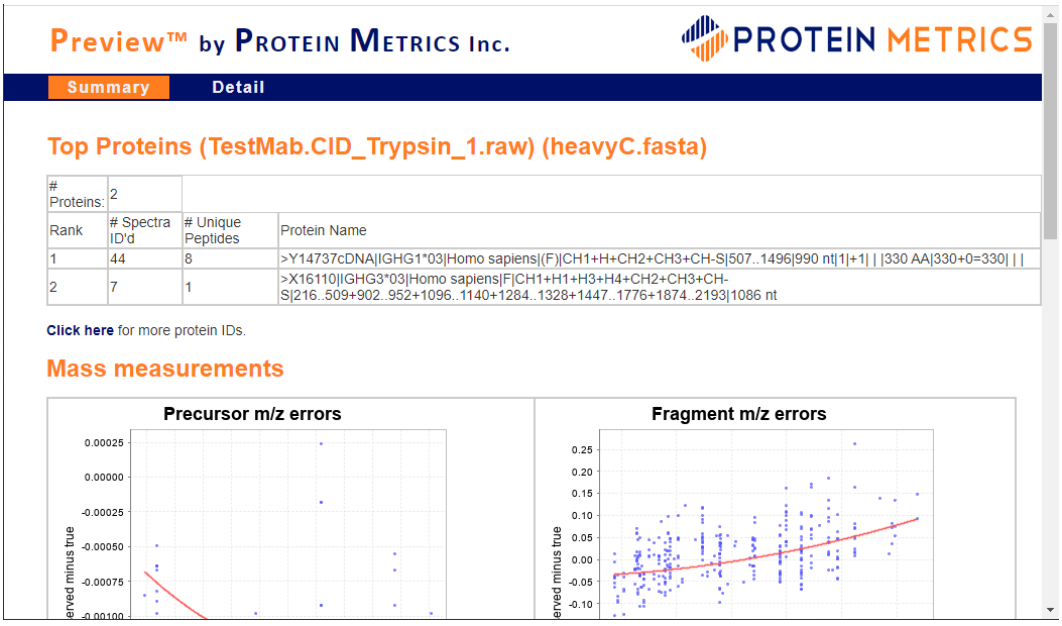


Figure 17: Preview Summary results as a web page

Click **Details** in the top bar to see detailed results:

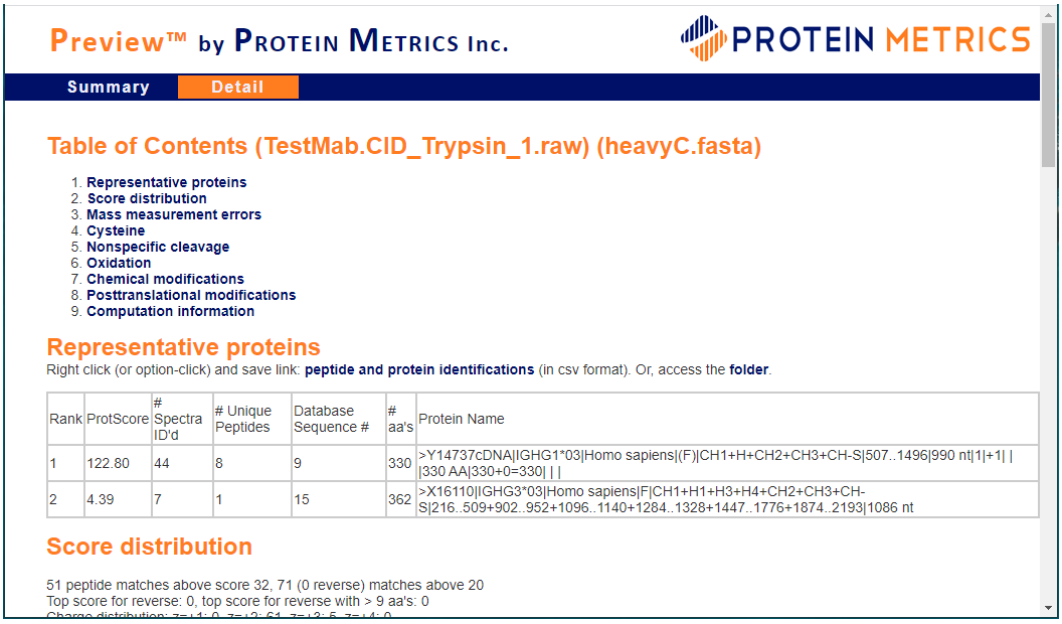



Figure 18: Preview Detail results

If **Launch Byonic** was checked before the run, Byonic opens, loaded with the parameters used to run Preview:

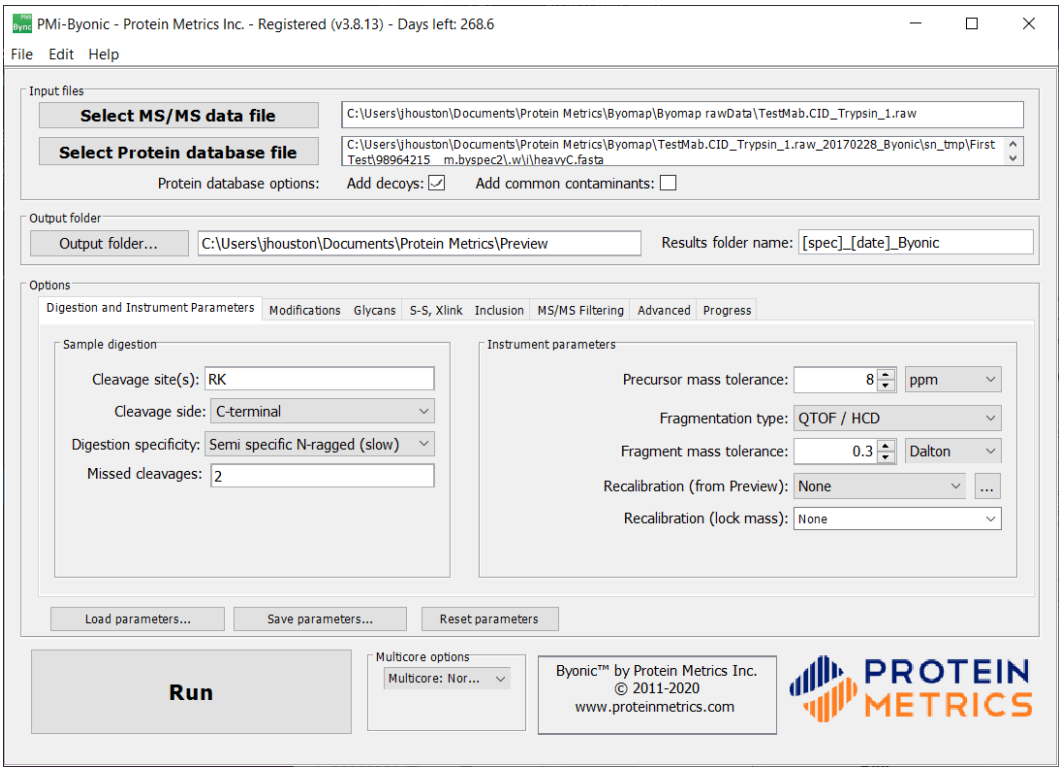



Figure 19: The Byonic application opened and loaded with Preview parameters

A copy of the Byonic parameter file (*.byparms) is also saved to the output file during the run. This can be used in future Byonic searches.

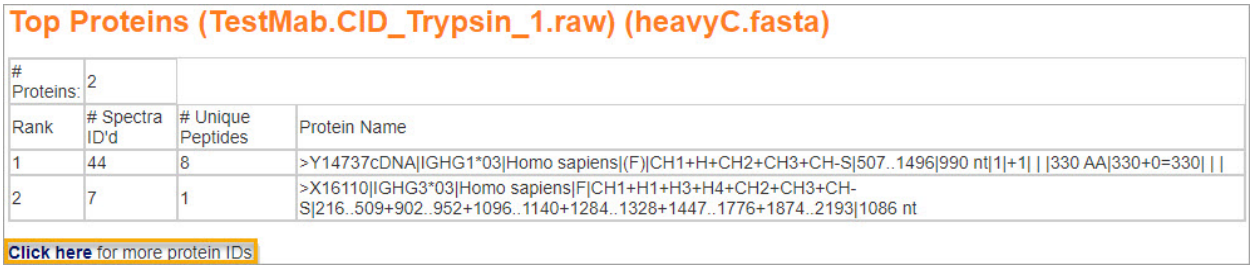
## Preview Summary Results

The Preview Summary results display a basic report of the run. Note that there will be fewer identifications than can be obtained using standard search engines. Preview is designed to sample data, not perform exhaustive searches.

The summary version contains six sections: **Top Proteins**, **Mass measurements**, **Digestion**, **Fixed modifications**, **Most common variable modifications** and **Detailed results**.

### Top Proteins

The **Top Proteins** section displays the best protein matches:



Top Proteins (TestMab.CID_Trypsin_1.raw) (heavyC.fasta)

| # Proteins: | 2 | | |
|---|---|---|---|
| Rank | # Spectra ID'd | # Unique Peptides | Protein Name |
| 1 | 44 | 8 | >Y14737cDNA\|IGHG1*03\|Homo sapiens\|(F)\|CH1+H+CH2+CH3+CH-S\|507..1496\|990 nt\|1\|+1\| \| \|330 AA\|330+0=330\| \| |
| 2 | 7 | 1 | >X16110\|IGHG3*03\|Homo sapiens\|F\|CH1+H1+H3+H4+CH2+CH3+CH-S\|216..509+902..952+1096..1140+1284..1328+1447..1776+1874..2193\|1086 nt |

Click here for more protein IDs

Figure 20: Summary results - Top Proteins

Preview ranks proteins using a function based on the number of unique peptides, the peptide matching scores, among other considerations. In the above example, only the top protein is surely in the sample, because it is the only protein with more than one distinct peptide. A few extra proteins do not hurt Preview's assays, because they will provide few high-scoring identifications. Preview automatically adds matched decoy peptides for all searches, and it uses the decoys to estimate and correct for false discoveries and to set the score thresholds for accepting identifications. There is no need to add decoys to the protein database.

To see the full list of identified proteins, click **Click here (**for more protein IDs). The detailed result page opens to the **Representative proteins** section. This contains the complete list of protein matches.

### Mass measurements

The **Mass measurements** section graphs m/z errors relative to the calibration curve:

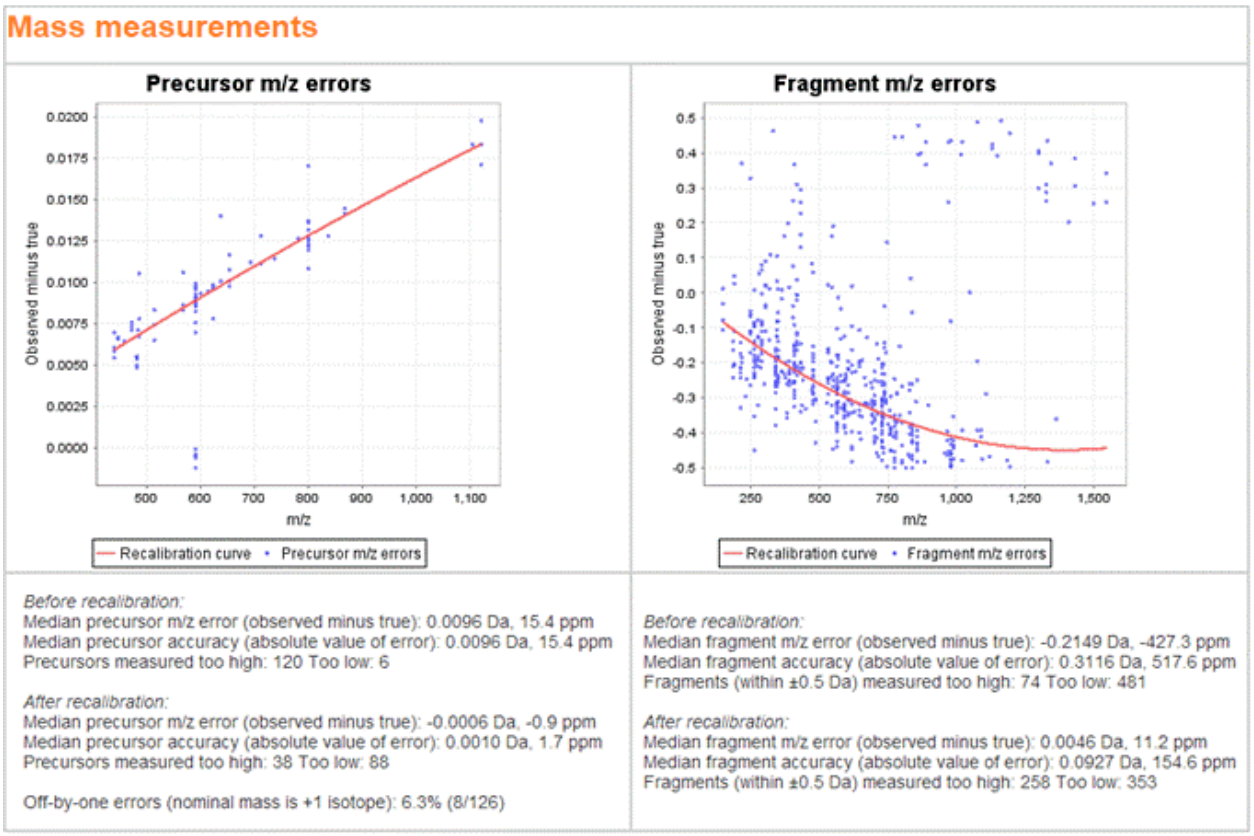



Figure 21: Summary results - Mass measurements

Below each graph are displayed the statistics for m/z error, both before and after recalibration.

The plots of precursor and fragment measurement errors reveal something interesting in the above example: the measurements benefit from recalibration. The precursor measurements ran about 15 ppm too high, which is a fairly large error for Orbitrap measurements. The fragment measurements ran about 0.3 Da too low, which is a fairly large error for LTQ measurements. Preview includes a built-in recalibration, which generates a new spectrum file (save to the specified output folder as Spectrum.identifications.csv). For this data set, recalibration improves the precursor errors by about 7-fold to a median accuracy (absolute value) of about 2 ppm and the fragment errors by about 3-fold to a median accuracy of about 0.1 Da.

Note that a median precursor accuracy of 2 ppm does not mean that the user should specify 2 ppm tolerance in a search engine. The median error is the typical error of an abundant ion, and the mass tolerance should be set to at least three times the typical error in order to catch all the valid identifications. A rule of thumb is to use five times the median error for the full Byonic search tolerance. In this case, we would choose a 10 ppm precursor tolerance and 0.5 Da fragment tolerance.

## Digestion

The **Digestion** section displays the digestion specificity and cleavage results:

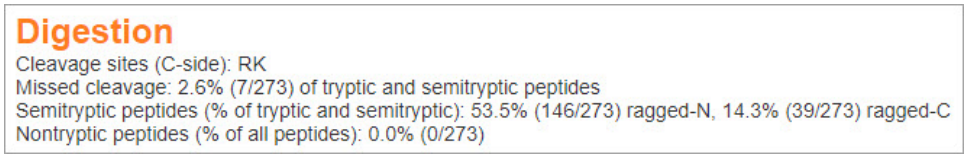



Figure 22: Summary results – Digestion

## Fixed modifications

The **Fixed modifications** section lists the fixed modifications found:




Figure 23: Summary results - Fixed modifications

In this example, no fixed modifications were pre-selected; only the modification found programmatically for cysteine is shown.

## Most common variable modifications

The **Most common variable modifications** section lists the occurrences of variable modifications:

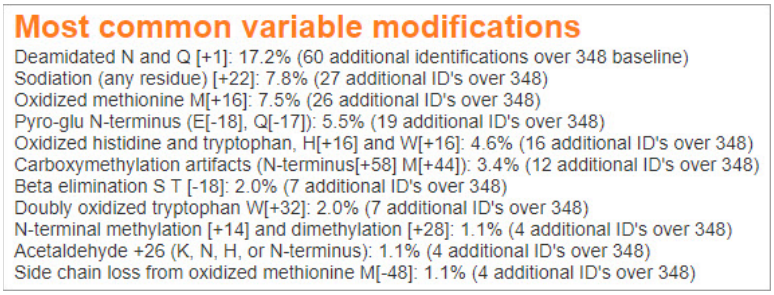



Figure 24: Summary results - Most common variable modifications

The list of the most common variable modifications helps the user choose the modifications to enable for a full search, based on prevalence, biological importance and search time. The user would probably want to enable all of the most common modifications in the Byonic search..

## Detailed results

**Detailed results** section displays links for accessing additional information about the results:

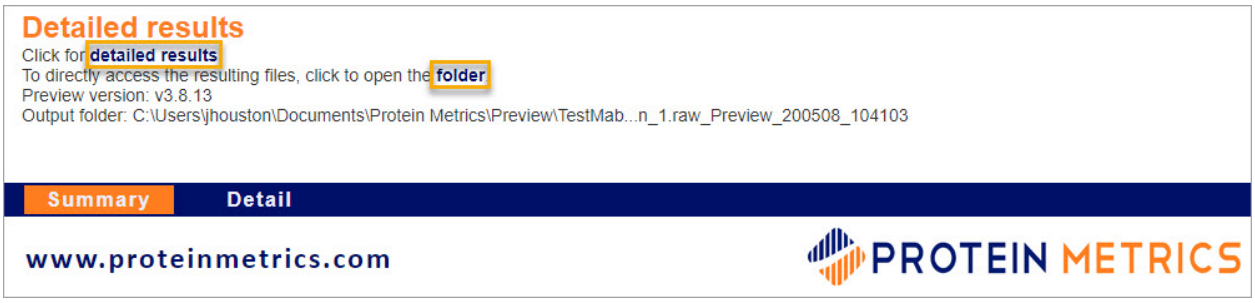



Figure 25: Summary results - Detailed results

- To go to the Detail results page, click **detailed results**. Alternatively, click **Detail** on the bottom bar.
- To open the output folder containing the results, click **folder**. The output folder is opened in the web browser:

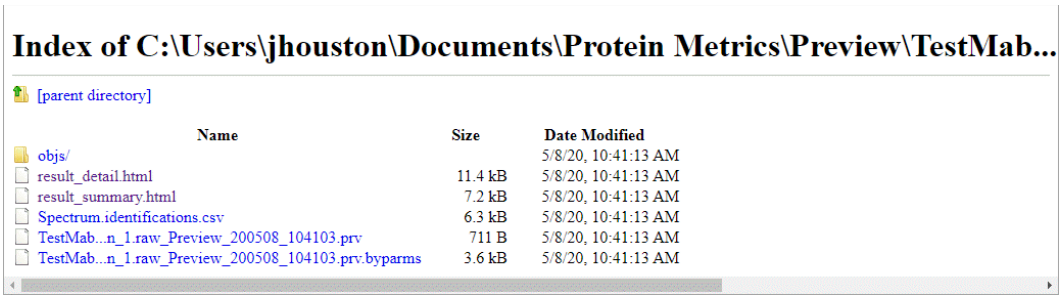



Figure 26: Output folder index

- **Preview version** may be useful when reporting issues.
- **Output folder** displays the path to the output folder. The folder is named for the source data file and the date. To open the folder in Windows Explorer, copy and paste the path into the top navigation cell in Explorer. The output folder contains the following:
    - The files used to generate the web pages
    - The Preview parameter file (*.prv) used in the run
    - A download of the Representative proteins table (*.csv) from the Detailed results page
    - A Byonic parameter file (*.byparms) suggested for use in a follow-up Byonic search

## Preview Detailed Results

The Detailed results page, accessed by a click on **Detail** in the top bar, lists all of Preview's assays, including both positive and negative results. It also lists the size of the sample for each assay. The sample sizes let the user judge the statistical significance of the assay result.

The detailed version contains nine sections, summarized in the Table of Contents.

### Table of Contents

The **Table of Contents** lists the sections with links to those sections:

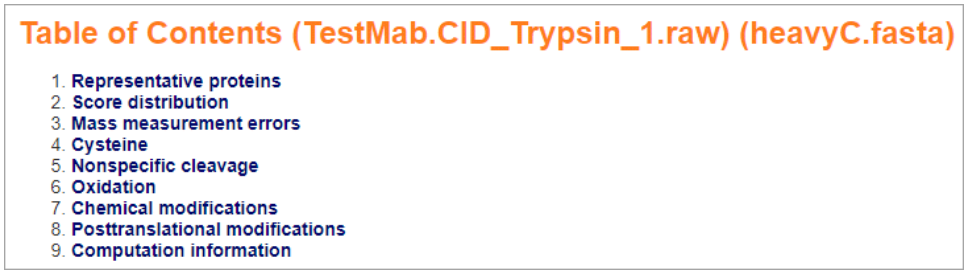
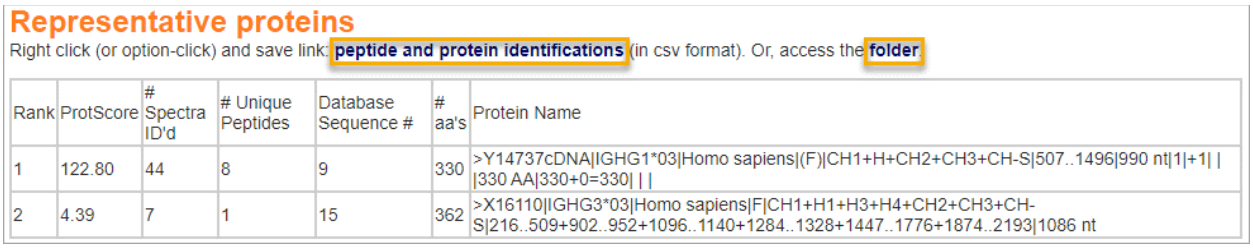



Figure 27: Detailed results - Table of Contents

### Representative Proteins

The **Representative Proteins** section displays a detailed table of all proteins found:



## Representative proteins

Right click (or option-click) and save link: peptide and protein identifications (in csv format). Or, access the folder:

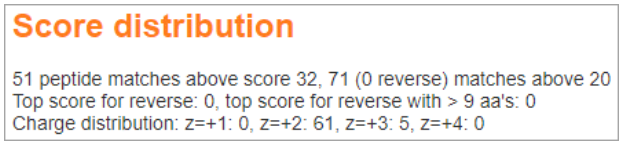
| Rank | ProtScore | # Spectra ID'd | # Unique Peptides | Database Sequence # | # aa's | Protein Name |
|---|---|---|---|---|---|---|
| 1 | 122.80 | 44 | 8 | 9 | 330 | >Y14737cDNA\|IGHG1*03\|Homo sapiens\|(F)\|CH1+H+CH2+CH3+CH-S\|507..1496\|990 nt\|1\|+1\| \| \|330 AA\|330+0=330\| \| \| |
| 2 | 4.39 | 7 | 1 | 15 | 362 | >X16110\|IGHG3*03\|Homo sapiens\|F\|CH1+H1+H3+H4+CH2+CH3+CH-S\|216..509+902..952+1096..1140+1284..1328+1447..1776+1874..2193\|1086 nt |

Figure 28: Detailed results - Representative Proteins

- To download the results, click **peptide and proteins identifications**. The results are incorrectly downloaded with the *.xls extension. To open this file in Excel, change the extension to *.csv. Alternatively, right-click the link and select save the page, rename the extension to *.csv in file name, save the type as **All Files**, and click **Save**. Note that a copy of the *.csv file is also saved to the output file during the run.
- To open the output folder containing the results, click **folder**. The output folder is opened in the web browser.

## Score distribution

The **Score distribution** summarizes the scores and charge distributions for the found proteins:



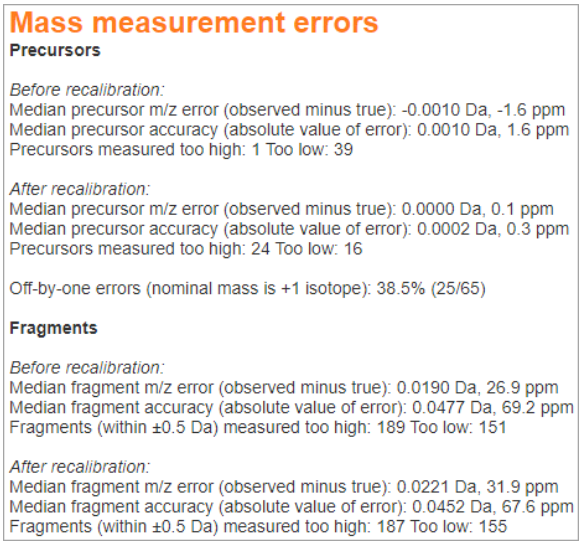
### Score distribution

51 peptide matches above score 32, 71 (0 reverse) matches above 20
Top score for reverse: 0, top score for reverse with > 9 aa's: 0
Charge distribution: z=+1: 0, z=+2: 61, z=+3: 5, z=+4: 0

Figure 29: Detailed results - Score distribution

## Mass measurement errors

The **Mass measurement errors** section displays the same the statistics for m/z errors displayed below the Mass measurements graphs in the Summary results:



### Mass measurement errors

**Precursors**

*Before recalibration:*
Median precursor m/z error (observed minus true): -0.0010 Da, -1.6 ppm
Median precursor accuracy (absolute value of error): 0.0010 Da, 1.6 ppm
Precursors measured too high: 1 Too low: 39

*After recalibration:*
Median precursor m/z error (observed minus true): 0.0000 Da, 0.1 ppm
Median precursor accuracy (absolute value of error): 0.0002 Da, 0.3 ppm
Precursors measured too high: 24 Too low: 16

Off-by-one errors (nominal mass is +1 isotope): 38.5% (25/65)

**Fragments**

*Before recalibration:*
Median fragment m/z error (observed minus true): 0.0190 Da, 26.9 ppm
Median fragment accuracy (absolute value of error): 0.0477 Da, 69.2 ppm
Fragments (within ±0.5 Da) measured too high: 189 Too low: 151

*After recalibration:*
Median fragment m/z error (observed minus true): 0.0221 Da, 31.9 ppm
Median fragment accuracy (absolute value of error): 0.0452 Da, 67.6 ppm
Fragments (within ±0.5 Da) measured too high: 187 Too low: 155

Figure 30: Detailed results – Mass measurement errors

## Cysteine

The **Cysteine** section lists the details determined for cysteine:

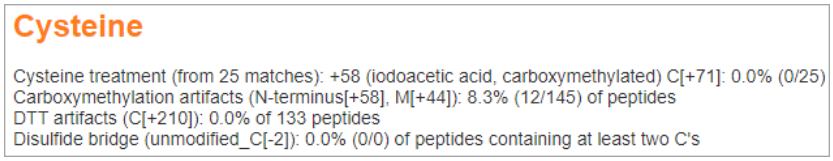



Figure 31: Detailed results - Cysteine

## Nonspecific cleavage

The **Nonspecific cleavage** section details all cleavage information:

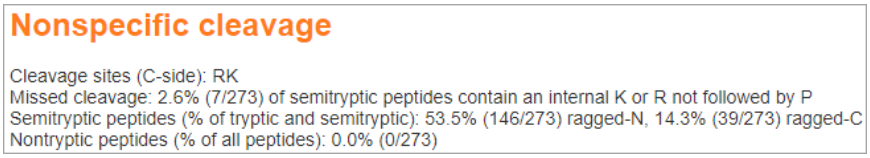



Figure 32: Detailed results – Nonspecific cleavage

## Oxidation

The **Oxidation** section details all oxidation information:

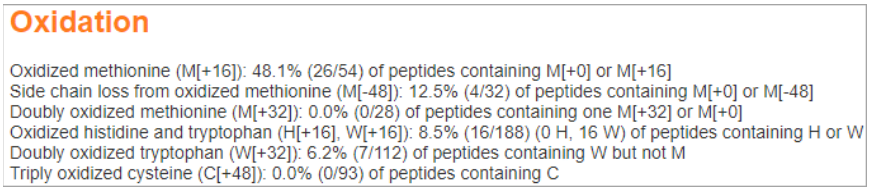



Figure 33: Detailed results - Oxidation

## Chemical modifications

The **Chemical modifications** section details all information for chemical modifications:

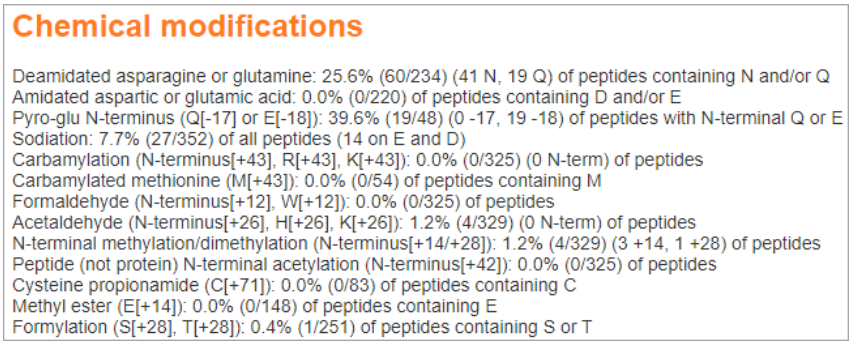



Figure 34: Detailed results – Chemical modifications

## Posttranslational modifications

The **Posttranslational modifications** section details all information for posttranslational modifications:

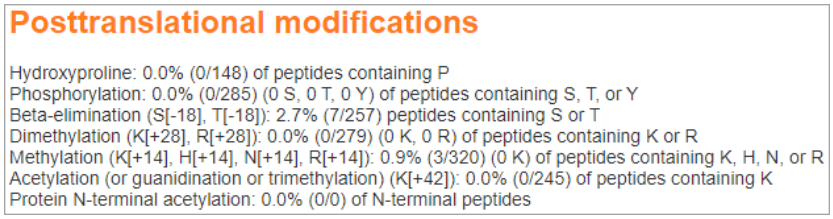



Figure 35: Detailed results – Posttranslational modifications

## Computational information

The **Computational information** section lists the parameters used in Preview and the output files generated in the run:




Figure 36: Detailed results – Computational information

- To download a parameter file (*.prv) with the parameters used in this run, click **parameters**. Note that a copy of the parameter file is also saved to the output file during the run.
- To return to the Summary results page, click **Summary** in the bottom bar**.**

Note: The Details page reports the rates of modification on eligible peptides, whereas the Summary page reports potential gains in the total number of identifications. Denominators in the percentages may also vary from search to search due to “second-order” effects such as multiply modified peptides and corrections for hits to decoys.

# Main Menu Bar

The main menu bar includes three items: **File**, **Edit**, and **Help**.

## File Menu

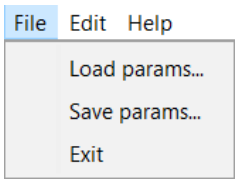



Figure 37: File menu

The **File** menu contains three items: **Load params**, **Save params**, and **Exit**. **Save params** saves a copy of all the parameters entered in the Preview search to a file with extension *.prv. **Load params** loads the parameter file saved from a previous session. The loaded parameters will replace existing parameters. These functions are the same as the **Load parameters** and **Save parameters** buttons at the bottom. **Exit** closes the application.

## Edit Menu

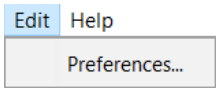



Figure 38: Edit menu

The **Edit** menu contains the Preferences option. **Preferences** allows the user to set the default folder where output files will be saved.

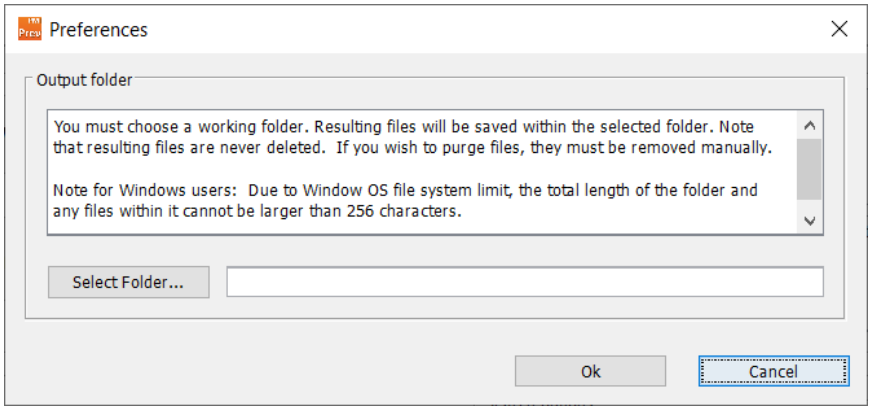



Figure 39: Edit > Preferences to set an output folder

If no output folder has been set before the search is run, this dialog will open so the user can set one.

## Help Menu

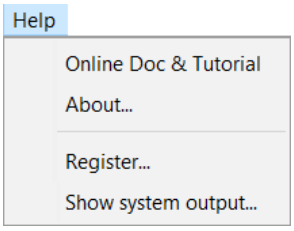



Figure 40: Help menu

The **Help** menu provides information about the software. **Online Doc & Tutorial** opens a video tutorial about Preview on www.proteinmetrics.com. The **About** menu displays the software version number and system information, which is useful when reporting issues. **Register** is used to activate Preview upon first use, along with other license actions. See the **PMI End-User License Manual** for more information. **Show system output** displays logs for the current Preview search; these are helpful for troubleshooting problems together with the Protein Metrics, Inc. staff via support@proteinmetrics.com.

# Preview FAQs

### Is Preview a search engine like Byonic, Mascot, or SEQUEST?

Not exactly. Preview samples the data, so it generally makes fewer identifications than a full search engine. On the other hand, it tests many more modifications and search options than any full search engine.

### How should I use Preview?

Run Preview before you run any other searches, so that you will know what type of full search will be most effective.

### Can I use Preview to recalibrate my m/z measurements?

Yes! If Preview makes sufficiently many identifications, say at least 20 precursors and at least 50 fragments, then you will generally get better results out of a full search with Preview's recalibrated spectrum file than with the original spectrum file unless the original calibration is extremely good. If you have enough identifications to avoid over-fitting (say 100 or more precursors), you can even run Preview's recalibrated spectrum file through Preview again for even more precise recalibration.

### How does Preview recalibrate m/z measurements?

Preview maps measured m/z values to recalibrated m/z values using quadratic curves – the red curves shown in the **Mass measurements** plots. Typically, calibration does not drift much over the course of an LC-ESI run, so that the same quadratic curve works for all spectra. Calibration can change from plate to plate with MALDI, however, so it is quite possible to see a lot of scatter in the m/z errors from a data set comprising many MALDI plates.

### If Preview reports median precursor error of 2 ppm, what should I set for the precursor tolerance in the full search?

The median error is the typical error for an abundant ion and at least 3 to 5 times smaller than the maximum error. Also check the number of "off-by-one" errors reported by Preview: even on high accuracy instruments, many precursor masses may reflect the mass of the first isotope peak rather than the monoisotopic mass.

### In the full search, should I enable all the modifications that Preview reports as Common variable modifications?

Not necessarily. Some full search engines do not support all the modifications supported by Preview. Some modifications are biologically uninteresting (for example, sodiation) and should only be enabled if they would contribute a significant number of additional identifications.

### How does Preview compute False Discovery Rate (FDR)?

Preview uses the target/decoy approach to FDR estimation: estimating the number of true identifications by the number of target identifications minus the number of decoy identifications. There is no need to add decoy proteins to the protein database; Preview does this automatically. Preview does not report FDR, but it uses FDR internally to decide which identifications to accept.

### How reliable are Preview's statistics?

Preview's statistics are especially good for "normal" shotgun proteomics, meaning digested multi-protein samples. Preview loses some reliability on very highly modified samples, in which many peptides carry more than one variable modification.

### How can I use Preview to improve my sample processing?

Preview reports on the amount of nonspecific digestion, m/z measurement errors, and sample preparation artifacts such as over- and under-alkylation, carbamylation, oxidation, sodiation, and deamidation. This type of information can provide valuable feedback.

### How should I read Preview's peptide and protein identifications?

This list (accessible from the Detail page) gives the highest scoring identification for each spectrum, so long as the score is high enough to be statistically significant. Too much should not be made of this list of identifications: remember that Preview samples the data, and does NOT perform a full search.

### How should I read Preview's wildcard search results?

Preview's definition of a wildcard is any mass shift on any one residue. Wildcard identifications are often approximate, with misplaced modifications, two modifications combined into one wildcard, two known modifications in a combination not considered by Preview's other searches, and so forth. On the other hand, these identifications, especially if they have scores over 60, are rarely completely wrong. A wildcard search will find polymorphisms, unanticipated modifications, and mystery mass shifts in almost any sample.

### Why do the Summary and Details statistics sometimes disagree?

The Summary page reports the overall gain to be achieved by enabling the modification, for example, 8.5% more identifications by allowing oxidized methionine for the BTK sample data. In contrast, the Details page reports the rate of modification, for example, 32.9% of peptides containing methionine contain at least one oxidized methionine. In other words, the Summary reports the "bottom line" of how many more identifications can be obtained by enabling the modification, while the Details page reports direct comparisons on specific, limited searches. For example, to assess the rate of oxidized methionine, Preview searches the spectra only against methionine-containing peptides and reports the results of the search on the Details page. After all searches have been done, Preview compiles the summary statistics by counting up all the identifications for all spectra.

Denominators in the percentages may also vary from search to search due to "second-order" effects such as multiply modified peptides and corrections for hits to decoys.

Preview's statistics can lose accuracy on extreme data sets, those in which a large percentage (say 30% or more) of the peptides carry more than one type of modification, for example, a data set that is both highly over-alkylated and highly oxidized.