Connecting Mascot Server 3.1 with Thermo Proteome Discoverer™

Version: 2025-01-15
Author: Matrix Science Ltd

Requirements

Mascot Server: version 3.1

Proteome Discoverer: versions 1.4, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 3.0, 3.1(*), 3.2

(*) The integration has been most thoroughly tested with PD 3.1. It is expected to work with all PD versions listed above. We will continue updating this document as we get more feedback from users.

Contents

Introduction

Thermo Proteome Discoverer™ includes a Mascot node, which provides an integrated connection with Mascot Server via HTTP/HTTPS. The Mascot node is compatible with Mascot Server 2.2 and later, including the latest versions, Mascot Server 3.0 and 3.1.

Mascot Server 3.0 (released September 2024) embeds MS²Rescore, which is a machine learning pipeline for rescoring database search results. The advanced machine learning (ML) features were not available through the Mascot node in Proteome Discoverer on release.

Mascot Server 3.1 is a patch release that includes a compatibility workaround to the HTTP/HTTPS API used by Proteome Discoverer. The API command used by the Mascot node for downloading the search result is client.pl?result_file_mime. The command sends the results in dat28 (MIME) format.

The data format and URL arguments are unchanged in Mascot Server 3.1. However, when the compatibility workaround is enabled, Mascot Server 3.1 automatically refines the results with machine learning, then formats the refined results in dat28 (MIME) format. This way, Proteome Discoverer gains access to predicted fragment intensities (MS²PIP) and predicted retention times (DeepLC) included in MS²Rescore.

The table below is an example of what is possible with Mascot Server 3.1. The full example is presented towards the end of the document.

Mascot Server	**Workflow**	**Protein Groups**	**Peptide Groups**	**Threshold**
3.0 (no refining)	Mascot →Target Decoy PSM Validator	4,635	21,507	Expect value: 0.87
3.1 (with MS2PIP:HCD2021 instrument)	Mascot →Target Decoy PSM Validator	5,898	31,141	Expect value (PEP): 0.1256

Setting up Mascot Server

If you have Mascot Server 3.0, please update to Mascot Server 3.1. The 3.1 patch release is free when you are under support. If your support contract has expired, please contact Matrix Science to renew support.

If you have Mascot Server 2.8 or earlier, you will need to purchase an update to Mascot Server 3.0. The version update includes 1 year of Premium Support, and the 3.1 patch release is free when you are under support.

After installing Mascot Server 3.1, there are four tasks:

1. Choose a suitable MS^{2}PIP model for your instrument and experiment.
2. Configure a new instrument definition that specifies the selected model.
3. Test the instrument definition in Mascot.
4. Check that the ClientResultFileMimeRefining option is enabled.
5. (Optional) Choose a suitable DeepLC model for your instrument and experiment.

1. Choosing a suitable MS^{2}PIP model

MS^{2}PIP is part of MS^{2}Rescore and provides fragment intensity predictions. The predictions are used as additional machine learning features for better separation between correct and incorrect matches.

Mascot Server ships with four MS^{2}PIP models that are suitable for Thermo instruments.

- **HCD2021**: Suitable when the acquisition uses HCD fragmentation. Use for qualitative studies as well as label-free quantitation.
- **CID**: Suitable when the acquisition uses CID fragmentation. Use for qualitative studies as well as label-free quantitation.
- **TMT**: Suitable for experiments using TMT or TMTpro labels.
- **iTRAQ**: Suitable for experiments using iTRAQ labels.

Selecting between HCD2021 and CID does not need to be guesswork. Go to your local Mascot home page and select Search Log. View a search you have recently submitted from Proteome Discoverer. In the format controls, enable refining with machine learning and choose an MS^{2}PIP model. After the report reloads, compare the FDR and match counts before and after refining to assess the model quality. Additional details are available in the machine learning quality report, which includes metrics for MS^{2}PIP model performance.

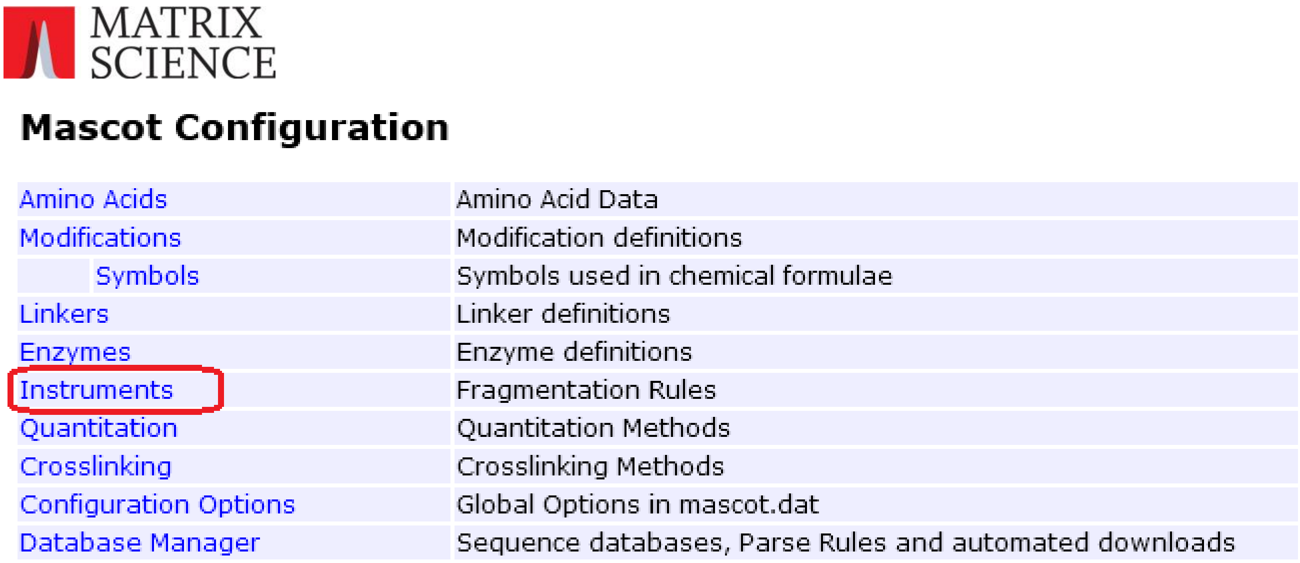
More information is available in our tutorial:

- https://www.matrixscience.com/blog/best-ms2pip-model-for-thermo-orbitrap.html
- Youtube: https://youtu.be/O7LSREqiW8o

2. Configuring a new instrument definition

The Mascot node in Proteome Discoverer 3.2 and earlier does not have a user interface for selecting an MS^{2}PIP model as a search parameter. Mascot Server now provides a workaround: the model can be configured as part of the instrument definition. When submitting the search from PD, select the new instrument to control whether machine learning should be used.

Go to your local Mascot home page and select Configuration Editor. Go to the Instruments panel.



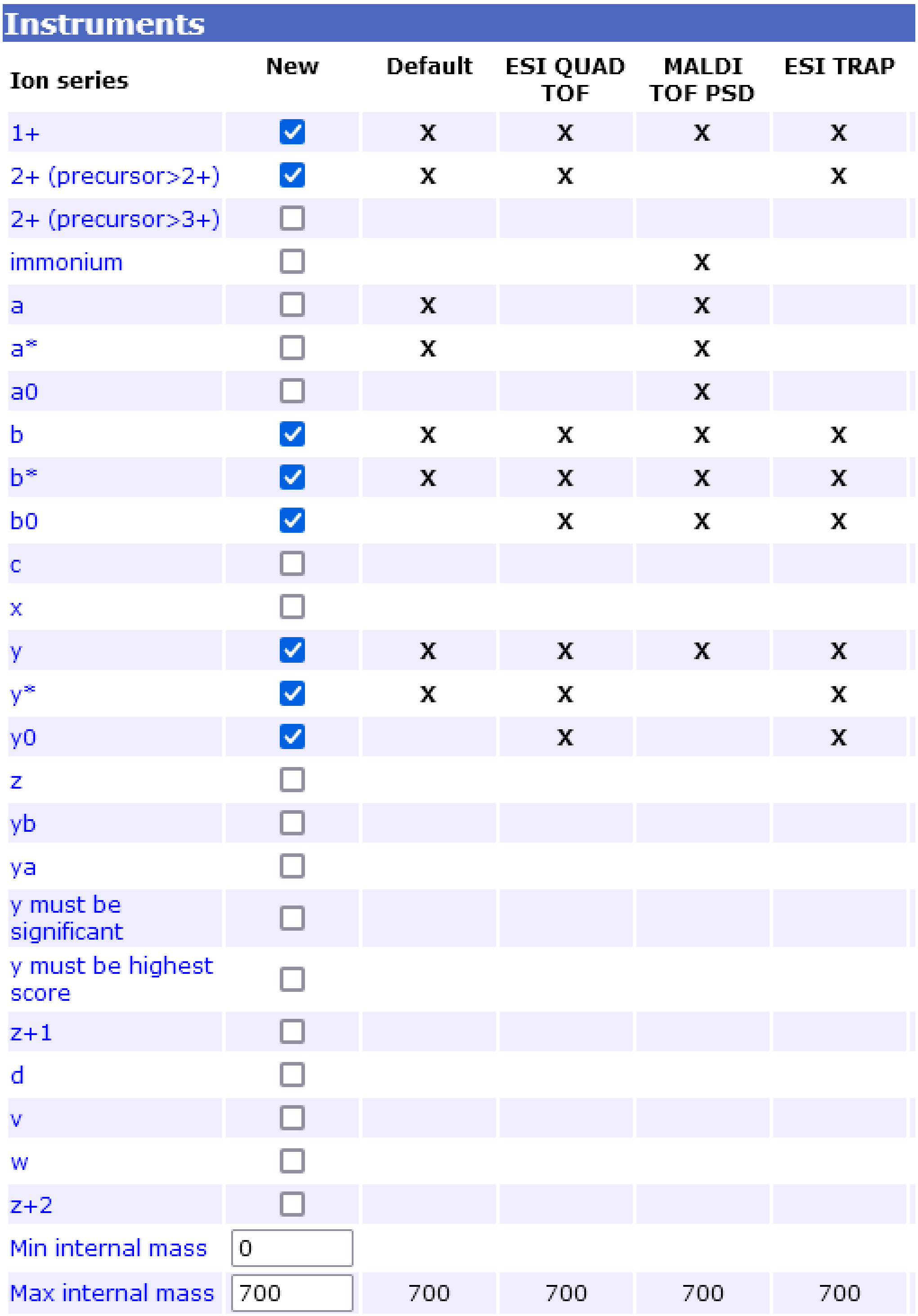
MATRIX SCIENCE

Mascot Configuration

Amino Acids	Amino Acid Data
Modifications	Modification definitions
Symbols	Symbols used in chemical formulae
Linkers	Linker definitions
Enzymes	Enzyme definitions
Instruments	Fragmentation Rules
Quantitation	Quantitation Methods
Crosslinking	Crosslinking Methods
Configuration Options	Global Options in mascot.dat
Database Manager	Sequence databases, Parse Rules and automated downloads

Create a new instrument definition. Give it a name that includes the MS^{2}PIP model name, so that it is easy to differentiate between instruments. For example, give it the name “MS2PIP:HCD2021”.

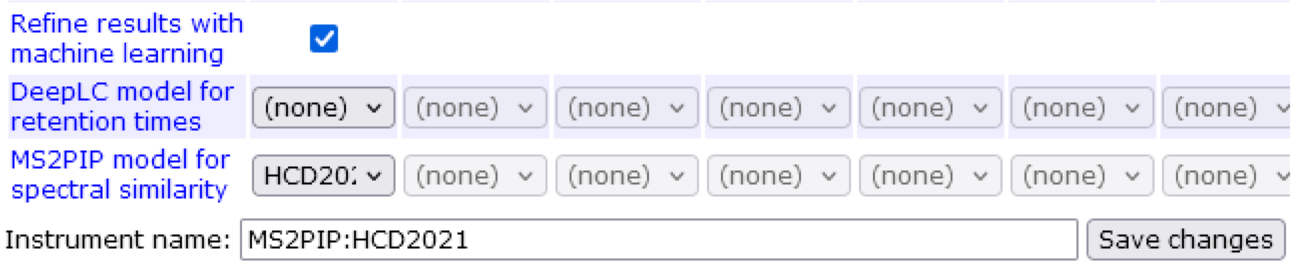
Select the fragmentation series based on an existing instrument. For example, if you regularly use the ESI-TRAP instrument when submitting searches from PD, select the same ion series:



Instruments

Ion series	New	Default	ESI QUAD TOF	MALDI TOF PSD	ESI TRAP
1+	☑	X	X	X	X
2+ (precursor>2+)	☑	X	X		X
2+ (precursor>3+)	☐				
immonium	☐			X	
a	☐	X		X	
a*	☐	X		X	
a0	☐			X	
b	☑	X	X	X	X
b*	☑	X	X	X	X
b0	☑		X	X	X
c	☐				
x	☐				
y	☑	X	X	X	X
y*	☑	X	X		X
y0	☑		X		X
z	☐				
yb	☐				
ya	☐				
y must be significant	☐				
y must be highest score	☐				
z+1	☐				
d	☐				
v	☐				
w	☐				
z+2	☐				
Min internal mass	0				
Max internal mass	700	700	700	700	700

Finally, enable refining with machine learning and choose the desired MS²PIP model:



Refine results with machine learning	☑						
DeepLC model for retention times	(none)	(none)	(none)	(none)	(none)	(none)	(none)
MS2PIP model for spectral similarity	HCD20:	(none)	(none)	(none)	(none)	(none)	(none)

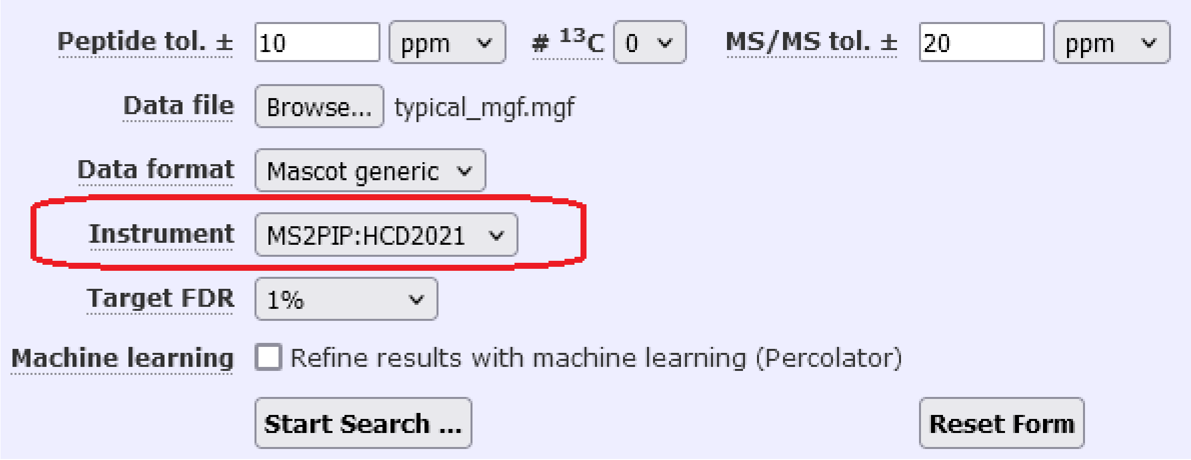
Instrument name: MS2PIP:HCD2021

Save. If you want to use several different MS^{2}PIP models, add a new instrument definition for each one.

3. Test the instrument definition in Mascot

Go to your local Mascot home page and Access Mascot Server. Select the MS/MS Ions Search.

Select a typical MGF file, and select SwissProt as the database. Choose the instrument you added (e.g. MS2PIP:HCD2021) and *disable* refining results with machine learning in the search form, as shown in the below screenshot:



Peptide tol. ± 10 ppm # ^{13}C 0 MS/MS tol. ± 20 ppm

Data file Browse... typical_mgf.mgf

Data format Mascot generic

Instrument MS2PIP:HCD2021

Target FDR 1%

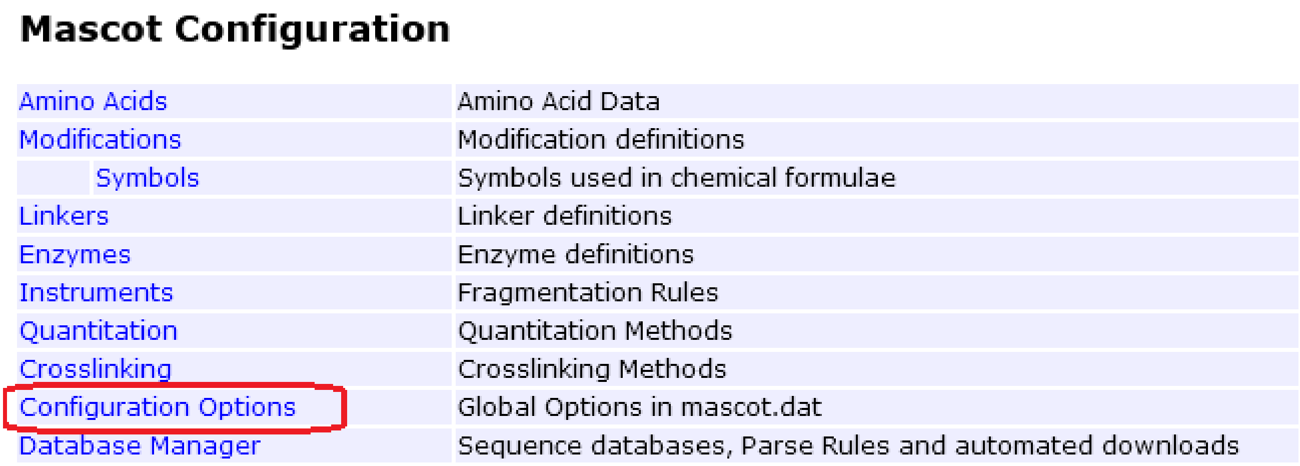
Machine learning ☐ Refine results with machine learning (Percolator)

Submit the search. When the search parameters are set this way, it simulates how the search is submitted from Proteome Discoverer. When the results report loads, it should default to refining the results with the MS^{2}PIP model selected in the instrument definition.

You can also perform this test by repeating an existing search from the Mascot search log.

4. Mascot option ClientResultFileMimeRefining

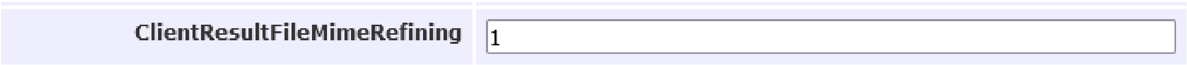
Go to your local Mascot home page and open Configuration Editor. Open Configuration Options.



Mascot Configuration

Amino Acids	Amino Acid Data
Modifications	Modification definitions
Symbols	Symbols used in chemical formulae
Linkers	Linker definitions
Enzymes	Enzyme definitions
Instruments	Fragmentation Rules
Quantitation	Quantitation Methods
Crosslinking	Crosslinking Methods
Configuration Options	Global Options in mascot.dat
Database Manager	Sequence databases, Parse Rules and automated downloads

Check that the option ClientResultFileMimeRefining is present and enabled (1). This option is enabled by default in a fresh Mascot Server 3.1 installation. Updating to Mascot Server 3.1 should also add and enable the option.



ClientResultFileMimeRefining 1

If the option is not present, add it and set its value to 1.

5. (Optional) Choosing a suitable DeepLC model

This step is optional but typically provides an additional boost to protein and peptide identifications.

DeepLC is part of MS²Rescore and provides retention time predictions. The predictions are used as additional machine learning features for better separation between correct and incorrect matches.

Once everything is working with the new integration using MS²PIP, you can follow the same procedure to select a suitable DeepLC model. Mascot ships with 20 DeepLC models, but it is sufficient to start with the model recommended by DeepLC developers, full_hc_PXD005573_mcp. More information is available in our tutorial:

- https://www.matrixscience.com/blog/tutorial-selecting-the-best-deeplc-model.html

After selecting a suitable model, add it to the instrument definition.

Setting up Proteome Discoverer

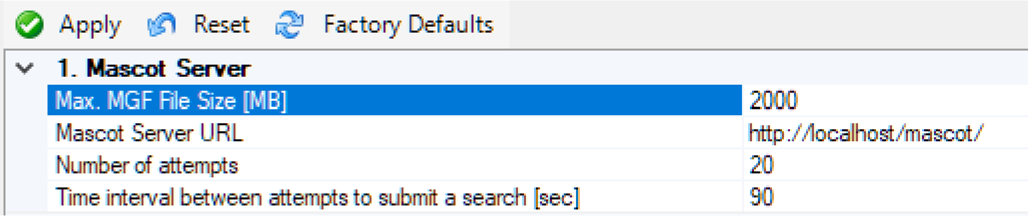
Mascot node configuration

The Mascot node in PD versions 1.4, 2.x, 3.0, 3.1 and 3.2 is fully compatible with Mascot Server 3.1.

The only change you should make is increase the Max. MGF File Size. If the peak list file is greater than Max. MGF File Size, then the Mascot node sends the peak lists to Mascot Server in chunks. This is intended to work around issues with web server limitations.

When refining with machine learning is enabled, it is very important that chunking is not used. Otherwise you will get suboptimal results, or refining may even fail if the chunk size is too small.

Increase the size to at least 2000MB:



1. Mascot Server

Max. MGF File Size [MB]	2000
Mascot Server URL	http://localhost/mascot/
Number of attempts	20
Time interval between attempts to submit a search [sec]	90

If Mascot Server is installed on Windows, then you are most likely using Microsoft IIS as the web server. IIS is limited to maximum upload size 2048 MB, so increasing the Max. MGF File Size beyond that will not help. If you peak lists are larger than 2GB, we recommend switching to the Apache web server.

Workflow

With Mascot Server 3.1, there are four possible processing workflows in Proteome Discoverer. Mascot is either connected to the Target Decoy PSM Validator node or the Percolator node; and the Mascot search is submitted either with an instrument enabling MS^{2}PIP, or an instrument that does not trigger refining with machine learning.

	Mascot → Target Decoy PSM Validator	**Mascot → Percolator node**
Mascot instrument name		
INSTRUMENT=ESI-TRAP	OK (case A)	OK (case B)
INSTRUMENT=MS2PIP:HCD2021	OK (case C)	Invalid (case D)

Case A: No machine learning is used.

Case B: Unrefined results are imported from Mascot into PD, then PD runs machine learning.

Case C: Mascot refines the results using machine learning, then they are imported into PD.

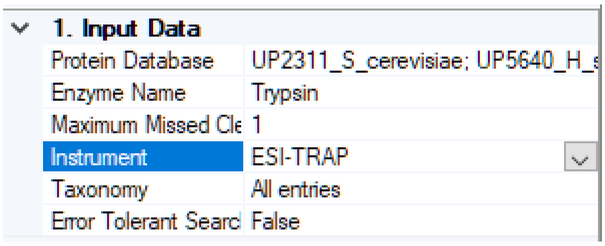
Case D: Invalid, please do not use.

You should never use the invalid workflow. In case D, Mascot runs Percolator to rescore the results after the database search. Then the refined results are imported into PD, and the Percolator node in PD runs another round of rescoring. These results are not statistically valid.

With Mascot Server 3.1, for simplicity, we recommend using workflows A and C, because it is easy to switch between them simply by changing the selected instrument in the Mascot node.

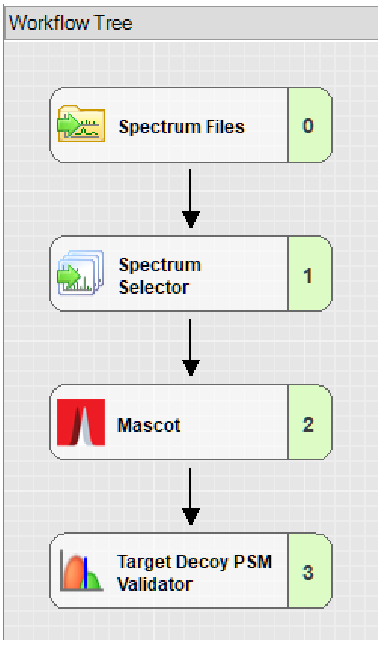
Case A:

The Mascot node is connected to the target-decoy validator, and the Mascot instrument is ESI-TRAP, or some other instrument where refining is not enabled.



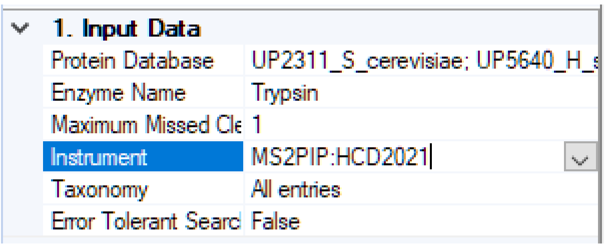
1. Input Data

Protein Database	UP2311_S_cerevisiae; UP5640_H_s
Enzyme Name	Trypsin
Maximum Missed Cle	1
Instrument	ESI-TRAP
Taxonomy	All entries
Error Tolerant Searc	False



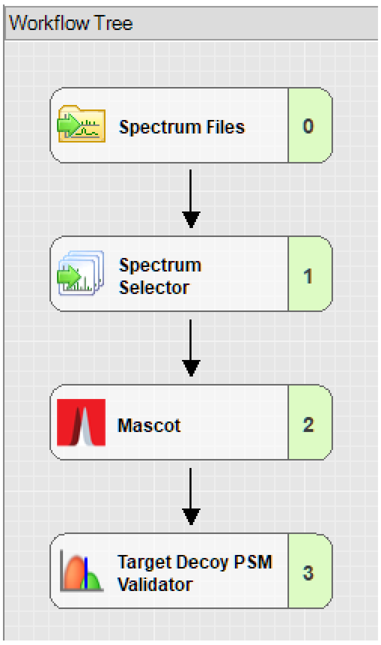
Case C:

The Mascot node is connected to the target-decoy validator, and the Mascot instrument is MS2PIP:HCD2021, or some other instrument that enables refining.



1. Input Data

Protein Database	UP2311_S_cerevisiae; UP5640_H_s
Enzyme Name	Trypsin
Maximum Missed Cle	1
Instrument	MS2PIP:HCD2021
Taxonomy	All entries
Error Tolerant Searc	False



Target Decoy PSM Validator settings

We recommend setting Target/Decoy Strategy to **Separate**. This is the correct strategy with Mascot, which runs separate target and decoy searches.

How the Mascot score is calculated and displayed in Proteome Discoverer

The Mascot node imports two key values from Mascot Server for each peptide-spectrum match (PSM): the Mascot ions score and qmatch. The identity threshold is calculated from qmatch and the significance threshold. Then, the node calculates the expect value from the score and the identity threshold. A PSM is significant if its expect value is below the significance threshold. Equivalently, a match is significant if its score is above the identity threshold.

When refining with machine learning is disabled, Proteome Discoverer displays the unmodified Mascot ions score and identity threshold. The identity threshold has a lower bound of 13, so any PSM with score below 13 can never be statistically significant, no matter how high you set the significance threshold.

When refining is enabled, Mascot sends a modified Mascot ions score and a modified qmatch to Proteome Discoverer. The modified score is -10*log10(PEP) + 7, where PEP is the posterior error probability estimated by Percolator. The modified qmatch is qmatch=100, which forces the identity threshold of all PSMs to 20 at the default significance threshold (effectively the same offset, 13+7).

If you view the same results in the Protein Family Summary report in Mascot Server, you'll notice that PSM score 13 is reported as score 20 in Proteome Discoverer. The reason for the offset +7 is to allow the target-decoy PSM validator to set the significance threshold to a value higher than 0.05. For example, to reach 1% FDR, the required PEP threshold may be 0.1256. This corresponds to the unmodified score -10*log10(PEP) = 9. However, because the identity threshold has a lower bound of 13, a PSM with score 9 (PEP 0.1256) could never become statistically significant. This is why Mascot sends the modified score -10*log10(PEP) + 7 = 16 to Proteome Discoverer and applies the same offset to the identity threshold. Now the target-decoy PSM validator is able to set a PEP threshold 0.1256 to reach 1% FDR.

Although the absolute value of the PSM score and identity threshold are different when viewing the results between the Protein Family Summary report in Mascot Server and the result viewer in Proteome Discoverer, the expect value of the PSM should be the same in both. For example, when refining with machine learning is enabled, a PSM with score 9 in Protein Family Summary has PEP 0.1256 and expect value 0.1256. The same PSM in Proteome Discoverer has score 16 and expect value 0.1256.

Please consult the Proteome Discoverer documentation about the difference between FDR, q-value and PEP.

Example: QC DDA run of human, yeast, *E. coli* mixture (PXD028735)

PRIDE project PXD028735 is raw data for A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics (Pyuvelde et al., Scientific Data, 9(126), 2022). The authors used six instruments and six different mixtures of yeast, *E. coli* and human proteins. Every sample was run with every instrument with DDA.

Download one of the QC replicates, such as LFQ_Orbitrap_DDA_QC_03.raw. This replicate was run on Thermo Orbitrap QE HF-X (Nano flow LC).

In Mascot Server 3.1, set up a new instrument MS2PIP:HCD2021, where refining is enabled and the MS^{2}PIP model HCD2021 is selected. The instructions for setting up the instrument are in the section above.

In Proteome Discoverer, create a new study. Use the consensus workflow template *ConsensusWF \ CWF_Basic.pdConsensusWF*. For the processing workflow, set up Mascot as shown in the above section for case C.

Use the following parameters for Mascot:

- Protein Database: SwissProt (alternatively, use human, yeast and *E. coli* Uniprot proteomes)
- Enzyme Name: Trypsin
- Maximum Missed Cleavages: 1
- Instrument: **MS2PIP:HCD2021**
- Taxonomy: All entries
- Error Tolerant Search: False
- Precursor Mass Tolerance: 10 ppm
- Fragment Mass Tolerance: 20 ppm
- 1. Dynamic Modification: Oxidation (M)
- 1. Static Modification: Carbamidomethyl (C)

Set the Target Decoy PSM Validator strategy to **Separate**. The rest of the processing settings (Spectrum Selector) and the consensus settings can be left at defaults.

Run the workflow. If the workflow succeeds, you should check a log file to confirm Mascot successfully refined the results and sent them to Proteome Discoverer. (There is currently no mechanism for Mascot Server to tell Proteome Discoverer that refining has been done. This requires an update to the Mascot node shipped with PD.)

1. Find the Mascot job number in the Proteome Discoverer the job queue. For example:
   ```
   Mascot      Info  Mascot result on server
   (filename=../data/20241215/F007074.msr)
   ```
2. Open the log file mascot\logs\workarounds\client_result_file_mime_refining.log on the Mascot Server hard disk. Confirm that the refined results were sent for this job, for example:

```
[result_file_mime][8412][../data/20241215/F007074.msr] Refining the
results.
[result_file_mime][8412][../data/20241215/F007074.msr] Refining succeeded.
[result_file_mime][8412][../data/20241215/F007074.msr] Preparing to combine
'../data/20241215/F007074.msr' with target and decoy pop files.
[result_file_mime][8412][../data/20241215/F007074.msr] About to run
combine_dat28_with_pop.pl
```

```
[result_file_mime][8412][../data/20241215/F007074.msr] Successfully created
MIME format file
..\data\cache\2024\12\wpjgy6beq7xy56bwbti7nokmmq\refined.dat.292179
[result_file_mime][8412][../data/20241215/F007074.msr] Dumped refined data
in MIME format to standard output. Done.
```

The below table summaries the protein and peptide counts when using Mascot Server 3.1 with the HCD2021 model, and Proteome Discoverer 3.1.

Mascot Server	**Workflow**	**Protein Groups**	**Peptide Groups**	**Threshold**
3.0 (no refining)	Mascot →Target Decoy PSM Validator	4,635	21,507	Expect value: 0.87
3.1 (with MS2PIP:HCD2021 instrument)	Mascot →Target Decoy PSM Validator	5,898	31,141	Expect value (PEP): 0.1256

The effective threshold can be found in the result viewer: In the PSMs tab, sort by Expect Value and find the largest value. With Mascot Server 3.0 and refining disabled, the target-decoy validator has to accept a lot of matches of dubious quality between expect values 0.05 and 0.87 in order to reach 1% FDR. With Mascot Server 3.1 and refining enabled, the target-decoy validator has a more sensible threshold. Not only is Proteome Discoverer detecting more matches, the matches are statistically more reliable.