Using machine learning with Mascot Server 3.1 and Thermo Proteome Discoverer™

Version: 2025-01-23
Author: Matrix Science Ltd

Introduction

Thermo Proteome Discoverer™ includes a Mascot node, which provides integration with Mascot Server via HTTP/HTTPS. The Mascot node is compatible with Mascot Server 2.2 and later.

With Mascot Server 3.1, you can automatically refine results with machine learning and import the refined results into Proteome Discoverer. Machine learning is optional and disabled by default. If you do not wish to enable it, then you do not need to follow the instructions below, as Mascot Server 3.1 works with Proteome Discoverer without extra steps.

If you enable machine learning, the table below illustrates the expected improvement, using a Thermo Orbitrap QE HF-X raw file from PRIDE project PXD028735.

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	34,141	Expect value (PEP): 0.1256

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 2.4 and 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.

Three steps are needed:

1. Configure a new instrument definition in Mascot
2. Increase the Max MGF Size in the Mascot node configuration
3. Set up a processing workflow using the new instrument definition

MATRIX SCIENCE

1. Configuring a new instrument definition in Mascot

The Mascot node in Proteome Discoverer 3.2 and earlier does not have a user interface for selecting an MS²PIP model. Instead, the model can be configured as part of the instrument definition.

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

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. 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	☐				

Enable refining with machine learning and choose the desired model. Give it a name that includes the model's name, so that it is easy to differentiate between instruments.

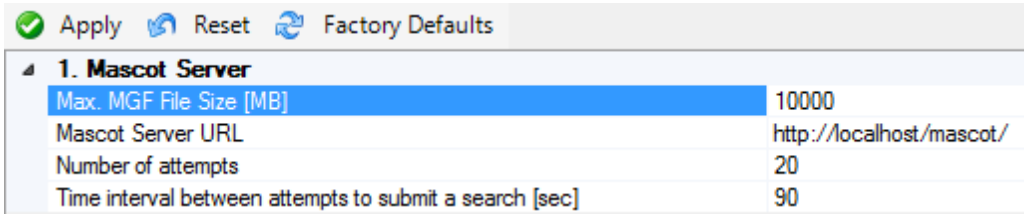
Mascot Server ships with four models that are suitable for Thermo instruments. If you want to use several different MS²PIP models, add a new instrument definition for each one.

- **HCD2021**: HCD fragmentation. Use for qualitative studies as well as label-free quantitation.
- **CID**: CID fragmentation. Use for qualitative studies as well as label-free quantitation.
- **TMT**: Experiments using TMT or TMTpro labels.
- **iTRAQ**: Experiments using iTRAQ labels.

2. Mascot node configuration

The Mascot node sends the peak lists to Mascot Server in chunks. 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.

In the node configuration, increase Max. MGF File Size to at least 10000MB:



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. If your peak lists are larger than that, you will get an upload error during Mascot search submission. An alternative to IIS is switching to the Apache web server on Windows or migrating Mascot Server to Linux.

The size of the MGF file is determined by the number of MS/MS scans. For reference, a 2h DDA run of a standard tryptic human sample on Thermo Orbitrap QE HF-X could record 110k MS/MS scans. The peak list is typically about 2kB per scan, so the MGF file would be around 210-230MB. An instrument with a higher scan rate will produce a larger MGF file.

3. Processing workflow

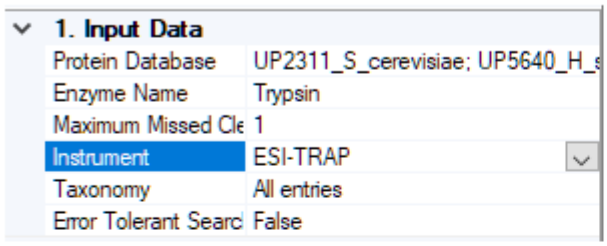
There are four possible processing workflows in Proteome Discoverer, tabulated below.

	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)

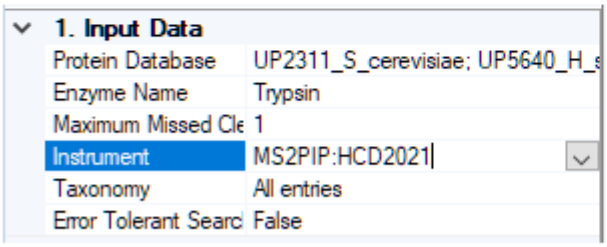
Case A and case C: For simplicity, we recommend using workflows A and C, because it is easy to switch between them simply by changing the selected instrument in the Mascot node. In case A, machine learning is not used; in case C, Mascot refines the results with machine learning.

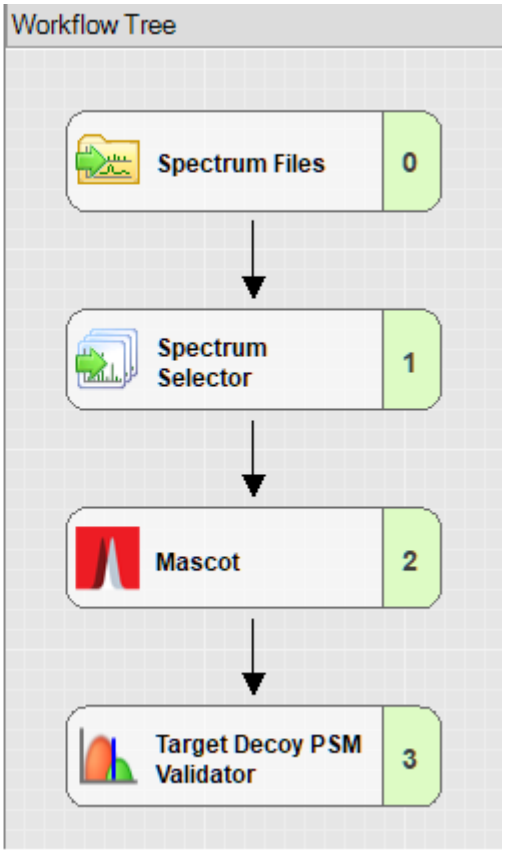
The Target Decoy PSM Validator node defaults to Concatenated strategy. We recommend setting the strategy to **Separate**, as Mascot runs separate target and decoy searches.

Case A:



Case C:





Case B: In case B, unrefined results are imported from Mascot into PD, then PD runs machine learning. No mechanism to choose an MS²PIP model.

Case D: You should never use this invalid workflow. In case D, Mascot refines results (using Percolator). Then the refined results are imported into PD, and PD runs Percolator again for another round of rescoring, without knowledge that the input data has already been rescored. These results are not statistically valid.