Customizable machine learning adapter improves LFQ in Mascot Distiller

Richard Jacob[1], Patrick Emery[2], Ville Koskinen[2]
[1]Matrix Science Inc, Boston, MA, USA
[2]Matrix Science Ltd, London, UK
richardj@matrixscience.com @matrixscience.bsky.social

COI: Ville Koskinen and Patrick Emery are directors and minority shareholders of Matrix Science Ltd.

How it works: A multi-file project is opened in Mascot Distiller. The raw data files, are processed and the peak lists sent to Mascot Server for searching. Once Mascot Server completes the searches it calls the adapter script, which reads the results and extracts peptide sequence match information. It passes this to the Machine Learning (ML) pipeline and runs it. The adapter inserts the predicted features from the ML into the Percolator input file (PIP) along with the core features from Mascot Server. Mascot Server runs Percolator for multidimensional analysis and rescoring creating a Percolator output file (POP). Mascot Server combines the posterior error probability (PEP) results with the initial search results and runs protein inference and passes them back to Mascot Distiller for quantitation. Mascot Server also generates a machine learning quality report. Mascot Distiller quantitates the results using XIC's.

Implementation: The initial implementation of the adapter script is with the MS²Rescore pipeline from the Martens Lab at the University of Ghent. Two ML components are included: DeepLC – which provides retention time prediction and includes 20 models covering different column types, gradient lengths and peptide properties; and MS²PIP - which provides spectral prediction and includes 13 models covering a range of fragmentation, instruments and peptide properties.

All the ML analysis is **run locally** on the Mascot Server and does not require a GPU.

By using the MS²Rescore pipeline better sensitivity is achieved at the same FDR. Depending on the data set increases of 20-30% more significant sequences is not unusual.

Developing Custom Adapters: Custom ML adapters for other pipelines can be written in any language supported by Mascot Parser (C#, C++, Java, Python, Perl)

Configuration: ML adapters are defined in a TOML file. This file specifies the program, interpreter (if required), whether it's visible in the user interface, command-line arguments, description, etc.

MS²Rescore Customization: It is possible to add custom DeepLC models but currently not custom MS²PIP models.

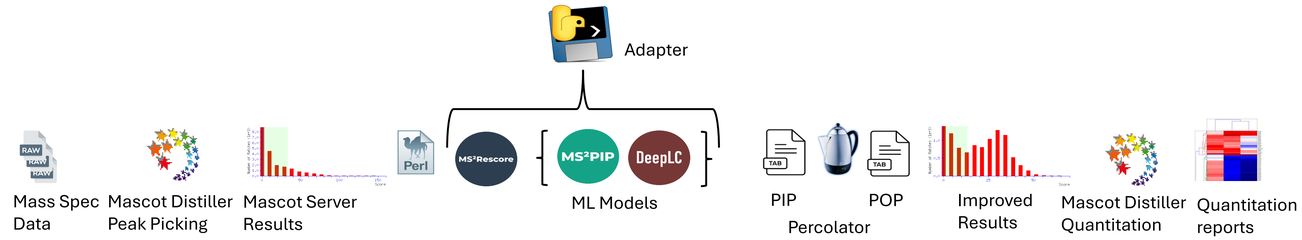
Example results
The encapsulated MS²Rescore pipeline was used with Mascot Distiller to quantitate a label free data set from the PRIDE project PXD028735 LFQ benchmark data set. Twelve Thermo Orbitrap QE HF-X raw files acquired using DDA methods were used with four each from Sample A, Sample B and the quality control sample.

Results were compared for A/B, A/QC, B/QC with and without MS²Rescore machine learning.

Mascot Server now has a **customizable adapter framework** allowing integration of third-party machine learning pipelines.

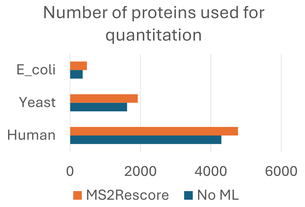
The initial implementation is the **MS²Rescore pipeline**, which leverages **DeepLC** retention time and **MS²PIP** spectral prediction models to refine search results.

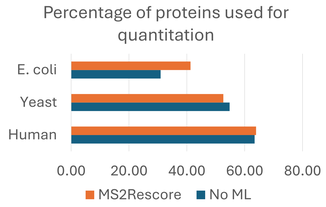
There are **significant improvements** in PSM and sequence matches after refining. These results carry through to more peptides and proteins being quantified at similar precision.



MS²Rescore leads to increased number of identifications

1% FDR	no ML	MS²Rescore	% Increase
PSMs	454113	549016	20.9
Sequences	39663	48428	22.1
Protein groups	5607	6049	7.9





Label free quantitation accuracy and precision are stable under machine learning

A/B	Expected Log2	No ML Median	Precision	MS²Rescore Median	Precision
Human	0.00	-0.05	0.08	-0.01	0.08
Yeast	1.00	1.02	0.14	0.99	0.19
E. coli	-2.00	-1.18	0.56	-1.29	0.65

A/QC	Expected Log2	No ML Median	Precision	MS²Rescore Median	Precision
Human	0.00	-0.09	0.07	-0.09	0.08
Yeast	0.42	0.38	0.08	0.36	0.09
E. coli	-1.32	-0.74	0.38	-0.85	0.48

B/QC	Expected Log2	No ML Median	Precision	MS²Rescore Median	Precision
Human	0.00	-0.05	0.10	-0.09	0.11
Yeast	-0.58	-0.60	0.14	-0.58	0.17
E. coli	0.68	0.59	0.07	0.46	0.17

Accuracy: median log2 protein ratio per species.
Precision: median absolute deviation (MAD) from the median log2 ratio.

MATRIX SCIENCE