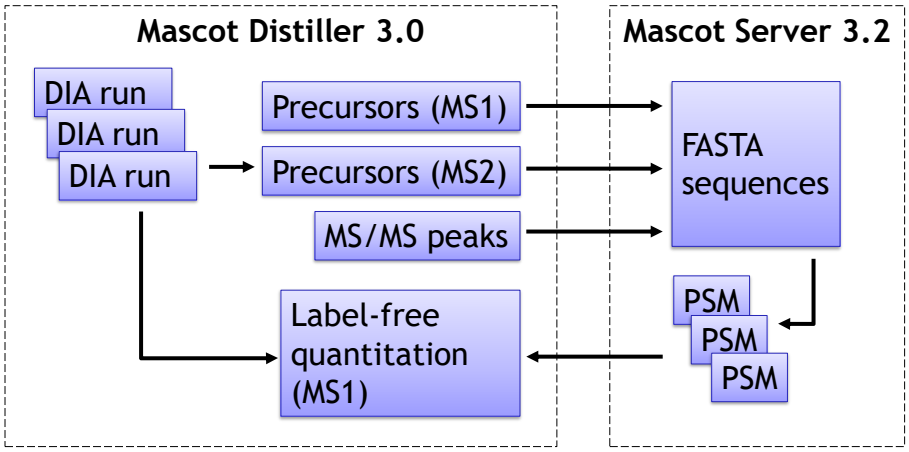
Mascot DIA:
Universal spectrum-centric approach

Ville Koskinen, Matrix Science Ltd

For quite a while now, we have been working on adding support for data independent acquisition into Mascot. We are planning to release what we call Mascot DIA later this summer, and this is a preview of what's coming up.

First of all, Mascot DIA is a fresh approach. It identifies peptides directly from the observed MS/MS spectra using a universal, spectrum-centric approach. I'll explain these terms shortly.

What is “Mascot DIA”?



Mascot DIA is composed of two parts: Mascot Distiller and Mascot Server. DIA support is implemented in the next version of Mascot Distiller, which is version 3.0 and currently in beta testing.

Here’s how it works:

- Mascot Distiller: Take DIA runs as input. Detect precursors from MS1 scans and (new!) from MS/MS scans. Detect fragment peaks from MS/MS scans. Submit the precursor masses and peak lists to Mascot Server.

- Mascot Server: Digest the protein sequence database into peptide sequences. Then, match all the detected precursor into the MS/MS spectrum, allowing multiple matches per spectrum.

Finally, Mascot Distiller downloads the search results and optionally runs label-free quantitation.

Maybe this looks familiar to you, because it sounds exactly like DDA processing. That's on purpose. The key differences are: precursor detection and how matches are

scored in the database search.

Spectrum-centric approach

- **Peptide-centric**
  - Start from a library of peptides (DDA runs or predicted from FASTA)
  - Is there evidence for the peptide in the observed data?
- **Spectrum-centric**
  - Start from the spectrum and peptide sequences (FASTA or spectral library)
  - Explain as many peaks in the observed spectrum as possible

Mascot DIA is a spectrum-centric approach, like Mascot DDA. Spectrum-centric and peptide-centric are technical terms that are mainly of interest to search engine developers, but I'll summarise the practical differences.

From a pragmatic point of view, a peptide-centric approach always requires a spectral library. The library could be made from DDA spectra or it could be made by predicting peak intensities using Prosit or MS2PIP or similar tool. But the spectral library is always there, even if you can't see it.

The peptide-centric approach starts from peptides in the library and tries to find evidence for them in the raw data.

In contrast, the spectrum-centric approach starts from the observed spectrum. The goal is to explain as many observed peaks as you can, in every mass spectrum.

The key consequence is, in a peptide-centric approach, you might identify a peptide using as few as 5 or just 3 fragments. In a spectrum-centric approach, a peptide is only identified if it has a confident sequence ladder in the mass spectrum.

Spectrum-centric approach

- **Peptide-centric, library-limited**
  - Spectronaut
  - DIA-NN
  - MSFragger-DIA
  - OpenSWATH
  - Skyline
  - EncyclopeDIA/Scaffold DIA
  - MaxDIA
- **Spectrum-centric, model-limited**
  - CHIMERYS
- **Spectrum-centric, unlimited**
  - NEW: Mascot DIA

Fröhlich et al., MCP 23(8):100800 (2024)

Peptide-centric could also be called the label library-limited approach. The library is made from DDA runs or predicted from protein sequences, but you can only identify what is in your library.

Here are some of the most commonly used DIA software tools categorised by their type. You'll notice the majority of software tools are peptide-centric.

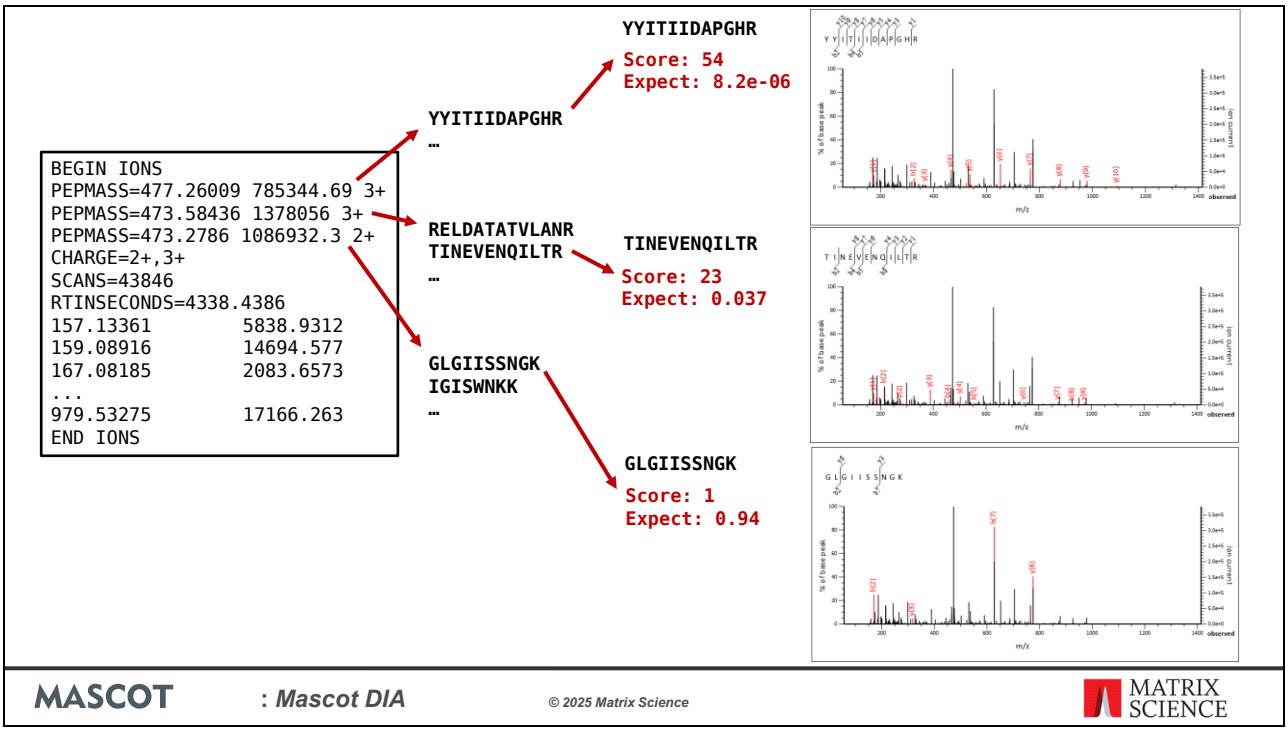
I've compiled this list from a recent review by Fröhlich et al.

(Fröhlich et al., MCP 23(8):100800 (2024), https://www.sciencedirect.com/science/article/pii/S1535947624000902)

There are not many spectrum-centric search engines. One of them is CHIMERYS. However, it also has limits in what it can find. CHIMERYS has a very smart machine learning model that predicts the fragmentation spectrum of a peptide, then it tries to sum the peaks of several peptides into the observed spectrum. However, the model is limited to whatever was in its training data. That means you can't choose an arbitrary enzyme or variable modifications that were not in the training set.

In contrast, Mascot DIA is also spectrum-centric, but it has no limits. You can choose any sequence database, any variable modifications and any enzyme, and search those against the observed data. Mascot doesn't need a library of predicted peptides or creating a DDA library. It identifies the peaks directly from the mass spectrum.

Mascot can make use of machine learning models, and I'll show an example in a bit.



Here is an example of an MS/MS spectrum in MGF format. It's a DIA spectrum with three precursor masses, specified here as PEPMASS lines. For DIA searches, Mascot requires that all the precursors are specified in the peak list.

During the database search, Mascot treats the precursors independently. For each one, it filters the sequence database for a list of plausible peptide sequences that are within precursor tolerance. In this case, it found one candidate for the first precursor, two for the second and two for the third.

Next, Mascot fragments the peptides *in silico* and counts the number of matching peaks. These are highlighted in red. Mascot matches each peptide independently to the same DIA spectrum. Notice that spectrum deconvolution is not necessary. All these sequences are matched to the observed, "convoluted" spectrum.

Finally, Mascot calculates a probability-based score. In this case, the top sequence got a very strong p-value, so it's unlikely to be a random match. The middle sequence has a weaker p-value, probably because its peaks are not very intense. And the last sequence got just four peak matches, even though it would explain one of those tall peaks. However, Mascot will never pretend that a match like that is significant. This is the major advantage of spectrum-centric searching.

Noise-resistant scoring (Mascot Server 3.2)

- **Peptide ions score (DDA):**
  - Select peaks by intensity
  - Count matches per ion series and residue coverage
  - Null hypothesis test: match is to an unrelated sequence
- **Peptide ions score (DIA):**
  - New probabilistic model using the same principles
  - Noise-resistant: ignore more noise and peaks from other precursors
- **Optimal isolation window width: 2m/z to 8m/z**
  - Wider ones are OK if spectral complexity is low

MASCOT : Mascot DIA MATRIX SCIENCE

The upcoming version of Mascot Server 3.2 has brand new probabilistic scoring for DIA. We call this noise-resistant scoring.

I don't have space to into great detail about scoring. Briefly, Mascot selects peaks by intensity and counts matches per ion series, as well as residue coverage. It does a null hypothesis test, where the null is: the match is to an unrelated sequence. The smaller the p-value, the less likely it's a random match.

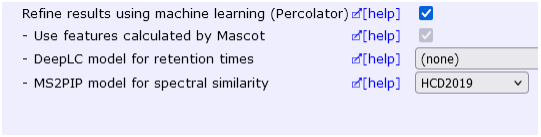
The noise-resistant scoring for DIA uses the same principles. It also selects peaks by intensity and counts the number of matches per ion series and residue coverage. However, the probability model is now able to ignore random noise peaks as well as peaks from other precursors. This means Mascot is able to match fragments from multiple precursors into the same spectrum.

One limitation of spectrum-centric searching is spectral complexity. Although the new noise-resistant scoring can go deeper into the peak list to identify a match, there is an issue when the spectral complexity is just too high. For standard tryptic digests, we have observed that 2 to 8 m/z is the optimal isolation window width. The best width is around 4 m/z. This is actually consistent with CHIMERYS, the other spectrum-centric tool, which also operates best in this range.

There isn’t anything that stops working if your isolation windows are wider. Wider windows can work if the spectral complexity is low. For example, if you have single-cell spectra, then 12 or even 16m/z could be fine. Or if you use variable windows, then use a narrow window at lower masses and wider window at higher masses where there are fewer precursors.

Using peak intensities and retention time

- **Noise-resistant scoring selects peaks by intensity**
  - Universal: independent of instrument characteristics
  - Provides high-quality matches for machine learning
  - But: wrong intensity pattern or wrong RT?
- **MS²Rescore integration:**
  - DeepLC for retention times
  - MS²PIP model for spectral similarity

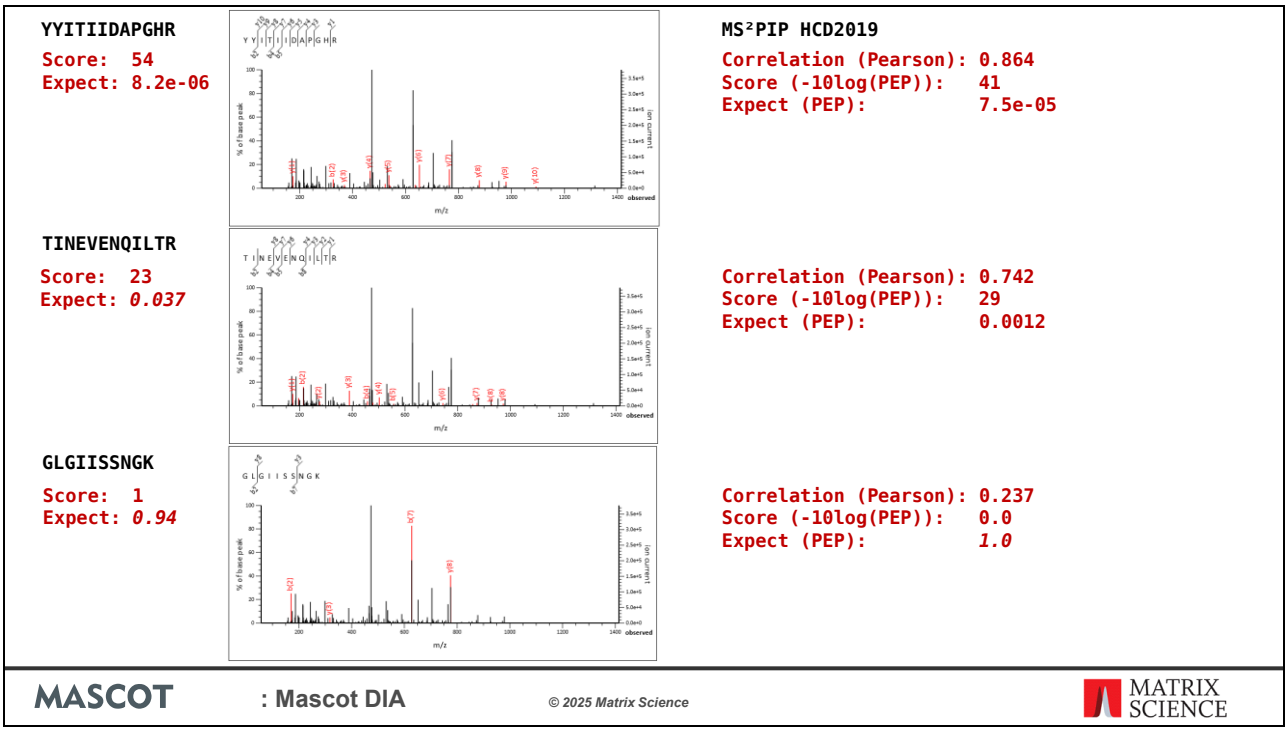


Now, Mascot scoring, including noise-resistant scoring, uses peak intensities to select signal from noise. It's a universal approach, because it doesn't depend on the instrument characteristics in any way. It also provides high-quality matches for machine learning.

One issue is, you might get a high score based on the fragments, but is it latching onto the wrong peaks? You might also get a high score for a sequence that elutes at the wrong time.

Thanks to the MS²Rescore integration, we can fix both problems. When you enable DeepLC, either in the search form or format controls, Mascot computes the difference between predicted and observed retention time. When you enable MS²PIP, Mascot uses the correlation between predicted and observed spectrum. These are fed to Percolator, which refines the results based on core features and the predicted features.

In this screenshot, I've selected the HCD2019 model, which is an excellent choice for Orbitrap instruments.



On the left, we have the example spectrum with its three precursors. These are the Mascot scores and expect values. The top one is statistically significant with very low expect value, but the middle one isn't, and the bottom one has too few fragment matches to be anything but a random match.

Let's enable the HCD2019 model. This has been trained on Orbitrap HCD spectra acquired in DDA runs.

Here are the correlations between the observed and predicted spectrum. The top one completely agrees with Mascot, high correlation, so it's nice to see it confirmed. The middle one is fairly high correlation too. After refining with machine learning, the posterior error probability (PEP) is quite low, 0.0012, so this match becomes significant.

The bottom match has very poor correlation with the predicted spectrum, so MS²PIP agrees with Mascot here as well. This is a random match before and after machine learning.

Rescoring with machine learning gets the best of both worlds. The spectrum-centric

matching explains as many peaks as possible, then correlation with predicted spectra includes the additional information about fragmentation patterns.

Example DIA MS/MS spectrum

```
BEGIN IONS
PEPMASS=477.26009 785344.69 3+
PEPMASS=473.58436 1378056 3+
PEPMASS=473.2786 1086932.3 2+
CHARGE=2+,3+
SCANS=43846
RTINSECONDS=4338.4386
157.13361         5838.9312
159.08916         14694.577
167.08185         2083.6573
...
979.53275         17166.263
END IONS
```

Requires accurate precursor determination!

Coming back to the example spectrum, you'll notice that Mascot relies on accurate precursor determination. It will only look for peptides if you give it the right precursor masses.

Precursor detection (Mascot Distiller 3.0)

- **From MS1 (survey) scan:**
  - Look for isotopic patterns within isolation window
  - Same for DDA and DIA
- **If failing to find precursors in survey scan: look in the MS/MS scan**

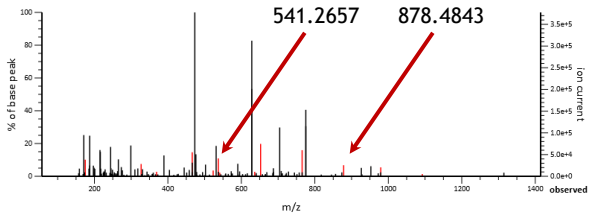
That's what the new version of Mascot Distiller does. When Distiller is processing a DIA run, it starts from the MS1 (survey) scan.

It divides the survey scan into isolation windows and select precursor masses within each window. The algorithm is the same for DDA and DIA.

In DIA mode, if Distiller fails to find precursors in the survey scan in the right isolation window, it switches to looking for precursor signal in the MS/MS scan. This is a new feature in Distiller 3.0.

Precursor detection (Mascot Distiller 3.0)

- **From MS/MS scan:**
  - Add masses of ion pairs to get MH+ plus proton; convert to 1+, 2+, …
  - Check for each charge state: is it in the isolation window?
  - Similar idea to Dančík et al., J Comp Biol, 6(3/4):327-342 (1999)



541.2657+878.4843 = 1419.75
➔473.5843, 3+
➔Within isolation window

Next, Distiller looks for precursor signal in the MS/MS scan. It uses a novel algorithm to infer precursor masses within the isolation window based on pairs of complementary ions.

Basically, add the masses of pairs of ions from the MS/MS spectrum. This gives you a precursor MH+ mass plus a proton. Convert it to Mr, then check whether it is in the isolation window. Do this for the relevant charge states (1+, 2+, so on). The principle is similar to a *de novo* algorithm proposed by Dancik et al. back in 1999.

Here's a quick example. We have two peaks at 541Da and 878Da. They sum to 1419.75. Take away a proton and convert to 3+ charge and you get 473Da. This is within the isolation window.

Precursor detection (Mascot Distiller 3.0)

- **From MS/MS scan:**
  - Add masses of ion pairs to get MH+ plus proton; convert to 1+, 2+, …
  - Check for each charge state: is it in the isolation window?
  - Similar idea to Dančík et al., J Comp Biol, 6(3/4):327-342 (1999)
- **Benefits:**
  - Data-driven (no library needed)
  - Agnostic to enzyme and variable mods
  - Precursor charge determination, even for centroided data
  - Mostly independent of fragmentation efficiency

This novel algorithm has several benefits.

- It's data-driven, so you don't need a library of precursor masses.

- It's agnostic to variable modifications. It doesn't even care whether the peptide is tryptic.

- It determines the precursor mass as well as charge, even with centroided data.

- It's mostly independent of fragmentation efficiency. As long as the peptide has fragmented enough to produce a few b and y ions, you should be able to find the precursor mass. And if the peptide hasn't fragmented, then you wouldn't be able to identify it anyway!

Mascot DIA: universal spectrum-centric approach

- **Data-driven approach: start from observed peaks**
- **Same workflow for DDA and DIA**
- **Mascot Distiller 3.0 (coming soon!)**
  - Initially with Thermo and SCIEX DIA support
  - Quantitation: any precursor-based, SILAC, dimethyl, LFQ
- **Mascot Server 3.2 (coming soon!)**
  - Noise-resistant probabilistic scoring
  - Any variable modifications; any enzyme
  - Isolation window width: best with 2-8m/z, can work for >8m/z

I've run out of time but I've shown you the key points about Mascot DIA. We've chosen spectrum-centric searching not because it's a gimmick, but because it yields certain benefits.

I've called Mascot DIA a *universal* approach. It's universal because:

- It's data driven. Start from the observed peaks. Mascot Distiller makes a list of precursor masses that were actually fragmented by the instrument. This is the opposite how all the other software tools work. It will work with peak lists from any instrument. It's not tied to any specific fragmentation patterns or peak intensity patterns.

- The same matching and scoring method is used with DDA and DIA. The search engine makes no assumptions about the data acquisition strategy. Same workflow for DDA and DIA, but it also means the isolation windows and scan ranges are flexible.

- Precursor detection and quantitation are implemented in the upcoming Mascot Distiller 3.0. It will initially support Thermo and SCIEX DIA files; we plan to add Bruker timsTOF support later.

- Existing quantitation workflows work the same with Mascot DDA and Mascot DIA. For example, SILAC or dimethyl labelling with DIA data is no different from DDA data. And obviously Distiller supports label-free quantitation.

- The upcoming Mascot Server 3.2 implements the new noise-resistant probabilistic scoring. You can use any variable modification or enzyme, as you don’t need a spectral library.

- Spectrum-centric matching works best when the isolation window width is between 2m/z and 8/mz, although it can work for larger windows too.