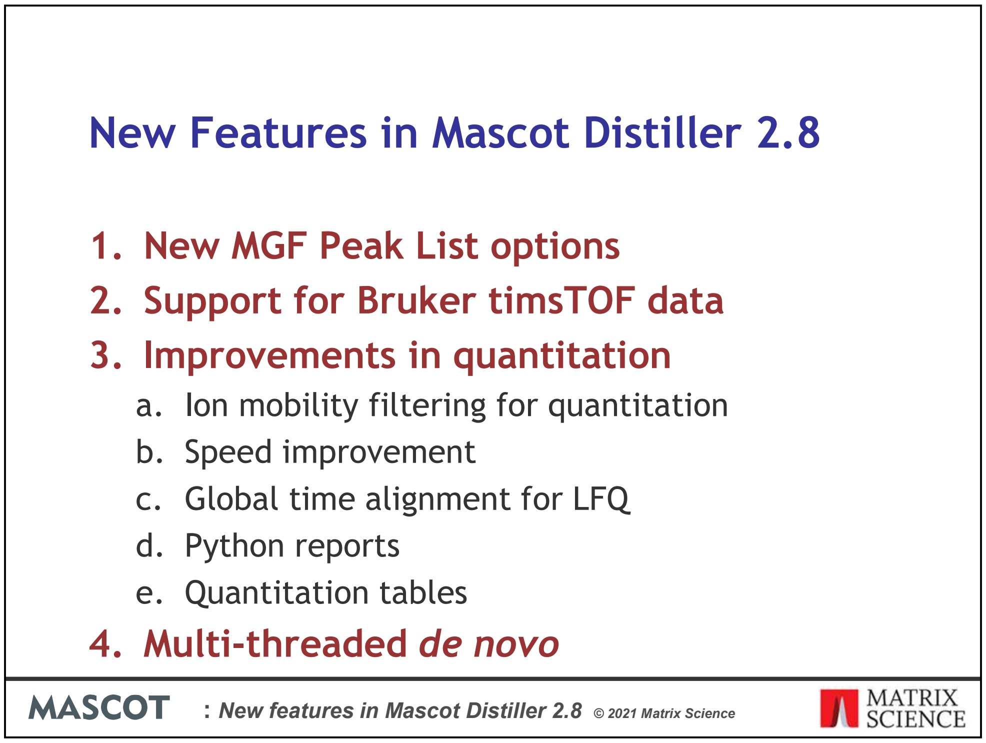
New features in Mascot Distiller 2.8

Patrick Emery



New Features in Mascot Distiller 2.8

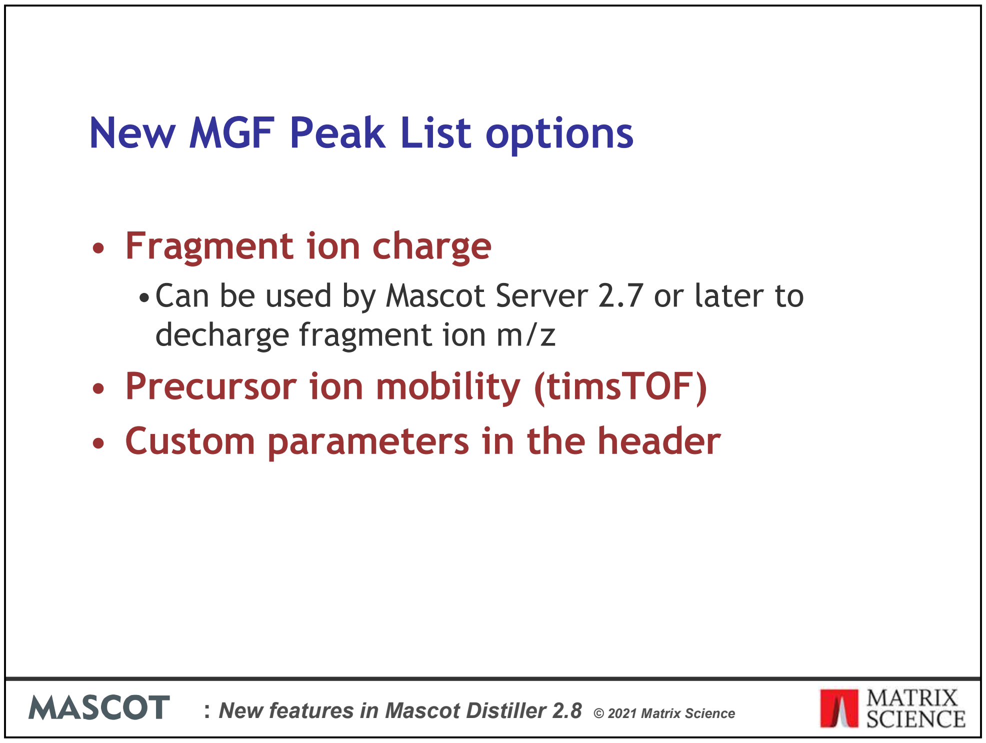
1. New MGF Peak List options
2. Support for Bruker timsTOF data
3. Improvements in quantitation
   a. Ion mobility filtering for quantitation
   b. Speed improvement
   c. Global time alignment for LFQ
   d. Python reports
   e. Quantitation tables
4. Multi-threaded *de novo*

MASCOT : *New features in Mascot Distiller 2.8* © 2021 Matrix Science MATRIX SCIENCE

There have been a number of improvements and new features added in Mascot Distiller 2.8.

The main area of improvement has been in precursor based quantitation – that's methods such as SILAC, precursor based label-free etc. These will be looked at in more detail in a separate talk.

In this presentation I'll be looking at:



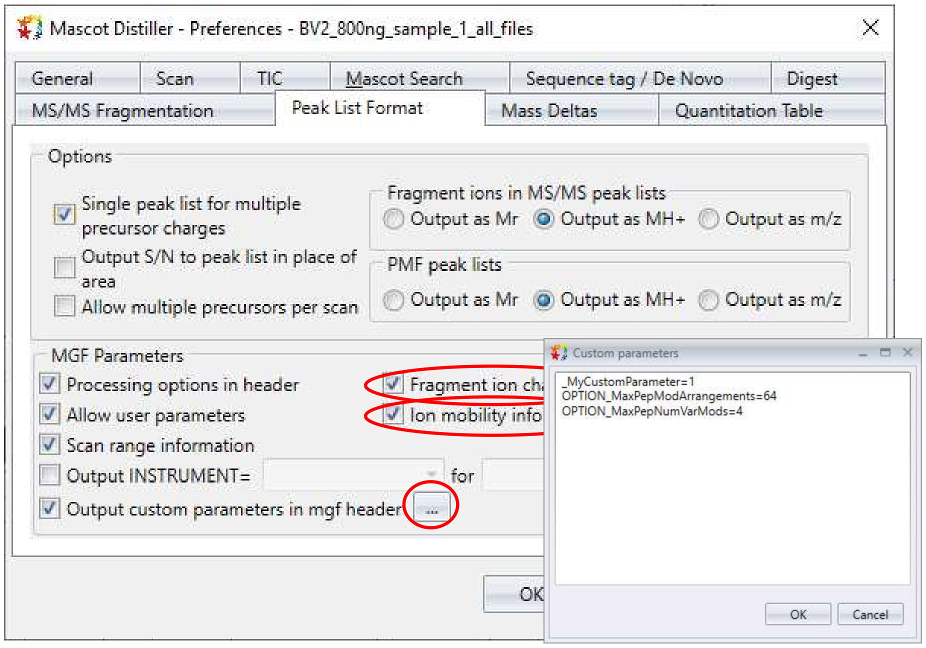
New MGF Peak List options

- Fragment ion charge
  - Can be used by Mascot Server 2.7 or later to decharge fragment ion m/z
- Precursor ion mobility (timsTOF)
- Custom parameters in the header

In Mascot Distiller 2.8 you can output a number of new features into any MGF peaklists generated. These include the charge of the individual fragment ions. Mascot Server 2.7 or later can use this information to decharge the MS/MS peaklist, which is important if you have charge states greater than 2+, and which can improve results for these types of data.

If you are processing timsTOF data, you'll want to include the precursor ion mobility in the peaklists. This requires Mascot Server 2.7 or later, and the values can then be used during precursor quantitation to filter the MS1 signal to improve quantitation

And you can add custom parameters to the MGF header. This can include search options, such as the options controlling variable modification permutation added in Mascot Server 2.7, or your own custom parameters which will be passed through to the results .dat file after searching. This can used for embedding useful metadata into the results.



Mascot Distiller - Preferences - BV2_800ng_sample_1_all_files

General | Scan | TIC | Mascot Search | Sequence tag / De Novo | Digest | MS/MS Fragmentation | Peak List Format | Mass Deltas | Quantitation Table

Options

- [✓] Single peak list for multiple precursor charges
- [ ] Output S/N to peak list in place of area
- [ ] Allow multiple precursors per scan

Fragment ions in MS/MS peak lists: Output as Mr [✓ Output as MH+] Output as m/z

PMF peak lists: Output as Mr [✓ Output as MH+] Output as m/z

MGF Parameters

- [✓] Processing options in header
- [✓] Allow user parameters
- [✓] Scan range information
- [ ] Output INSTRUMENT= for
- [✓] Output custom parameters in mgf header ...
- [✓] Fragment ion ch
- [✓] Ion mobility info

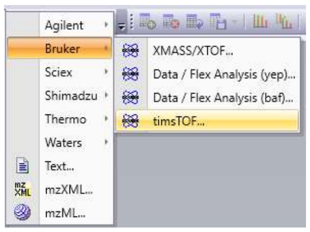
Custom parameters

```
_MyCustomParameter=1
OPTION_MaxPepModArrangements=64
OPTION_MaxPepNumVarMods=4
```

If you want to use any of these settings, go to Tools->Preferences in Mascot Distiller and select the “Peak List Format” tab. Here you’ll find checkboxes for enabling the output of fragment ion charges and precursor ion mobility values, and also where you can add a custom parameters to the MGF header.

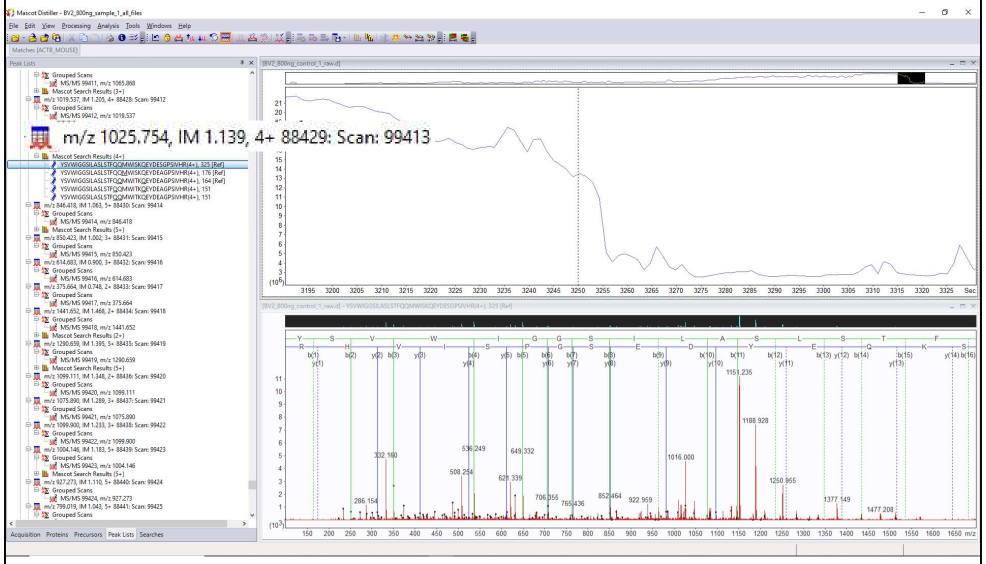
Bruker timsTOF support

- New instrument datatype
- MS/MS signals grouped by precursor ion mobility
- MS signal can be filtered by ion mobility for better quantitation (more later)



One of the key new features in Mascot Distiller 2.8 is support for the Bruker timsTOF data format. This is a new instrument data type, that you will find under the Bruker sub-menu when creating a new Project in Distiller.

MS/MS signals are grouped by precursor ion mobility. This allows us to filter the precursor MS signal to improve accuracy and reliability of quantitation – there will be more about that later in the presentation.



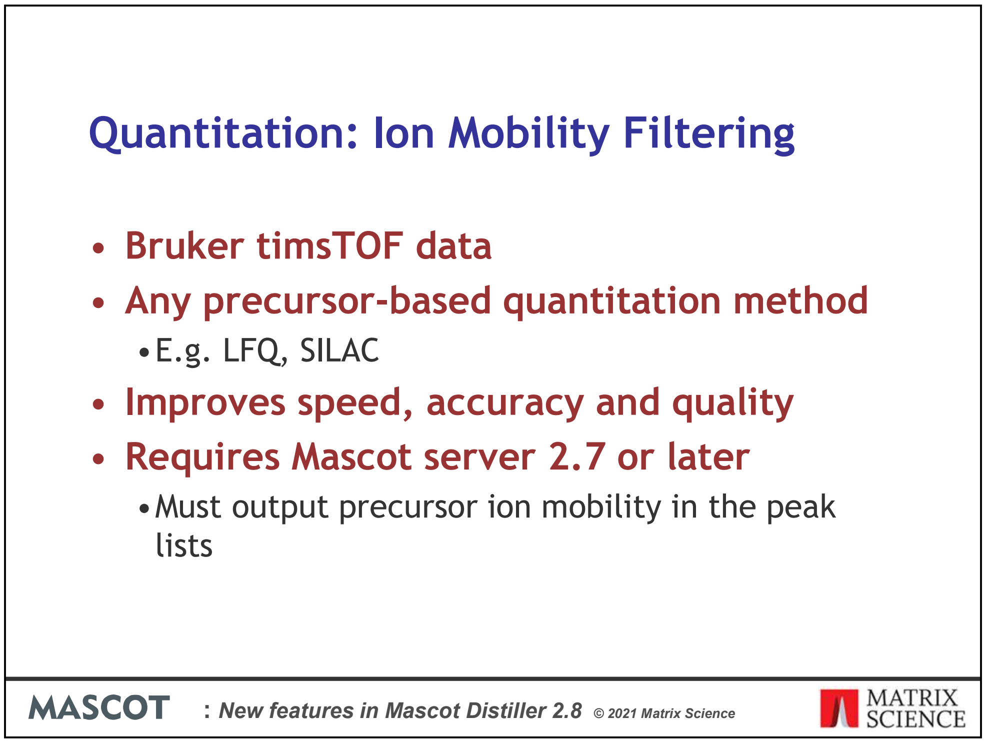
```
Mascot Distiller - BV2_800ng_sample_1_all_files
Matches [ACTB_MOUSE]
Peak Lists
    Grouped Scans
        MS/MS 99411, m/z 1065.868
    Mascot Search Results (3+)
m/z 1019.537, IM 1.205, 4+ 88428: Scan: 99412
    Grouped Scans
        MS/MS 99412, m/z 1019.537
m/z 1025.754, IM 1.139, 4+ 88429: Scan: 99413
    Mascot Search Results (4+)
        YSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHR(4+), 325 [Ref]
        YSVWIGGSILASLSTFQQMWISKQEYDEAGPSIVHR(4+), 176 [Ref]
        YSVWIGGSILASLSTFQQMWITKQEYDEAGPSIVHR(4+), 164 [Ref]
        YSVWIGGSILASLSTFQQMWITKQEYDEAGPSIVHR(4+), 151
        YSVWIGGSILASLSTFQQMWITKQEYDEAGPSIVHR(4+), 151
m/z 846.418, IM 1.063, 5+ 88430: Scan: 99414
    Grouped Scans
        MS/MS 99414, m/z 846.418
    Mascot Search Results (5+)
m/z 850.423, IM 1.002, 3+ 88431: Scan: 99415
    Grouped Scans
        MS/MS 99415, m/z 850.423
m/z 614.683, IM 0.900, 3+ 88432: Scan: 99416
    Grouped Scans
        MS/MS 99416, m/z 614.683
m/z 375.664, IM 0.748, 2+ 88433: Scan: 99417
    Grouped Scans
        MS/MS 99417, m/z 375.664
m/z 1441.652, IM 1.468, 2+ 88434: Scan: 99418
    Grouped Scans
        MS/MS 99418, m/z 1441.652
    Mascot Search Results (2+)
m/z 1290.659, IM 1.395, 5+ 88435: Scan: 99419
    Grouped Scans
        MS/MS 99419, m/z 1290.659
    Mascot Search Results (5+)
m/z 1099.111, IM 1.348, 2+ 88436: Scan: 99420
    Grouped Scans
        MS/MS 99420, m/z 1099.111
m/z 1075.890, IM 1.289, 3+ 88437: Scan: 99421
    Grouped Scans
        MS/MS 99421, m/z 1075.890
m/z 1099.900, IM 1.233, 3+ 88438: Scan: 99422
    Grouped Scans
        MS/MS 99422, m/z 1099.900
m/z 1004.146, IM 1.183, 5+ 88439: Scan: 99423
    Grouped Scans
        MS/MS 99423, m/z 1004.146
    Mascot Search Results (5+)
m/z 927.273, IM 1.110, 5+ 88440: Scan: 99424
    Grouped Scans
        MS/MS 99424, m/z 927.273
m/z 799.019, IM 1.043, 5+ 88441: Scan: 99425
    Grouped Scans
```

File Edit View Processing Analysis Tools Windows Help
Acquisition Proteins Precursors Peak Lists Searches

MASCOT : *New features in Mascot Distiller 2.8* © 2021 Matrix Science MATRIX SCIENCE

Once the timsTOF data are opened in Distiller, it looks very similar to other datatypes within the UI. The main difference you'll notice is that the precursor ion mobility value is shown for the peaklists.

Here we have a particularly impressive peptide match from a timsTOF dataset, processed with Mascot Distiller, with an ion score of 325.



Quantitation: Ion Mobility Filtering

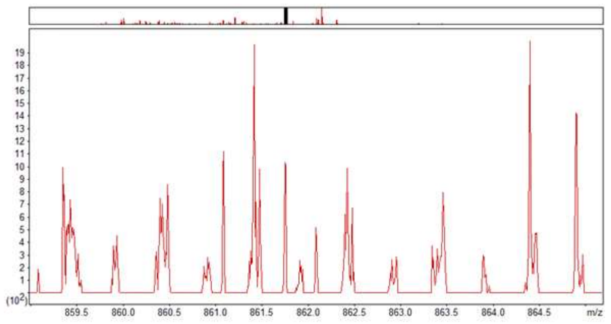
- Bruker timsTOF data
- Any precursor-based quantitation method
  - E.g. LFQ, SILAC
- Improves speed, accuracy and quality
- Requires Mascot server 2.7 or later
  - Must output precursor ion mobility in the peak lists

MASCOT : *New features in Mascot Distiller 2.8* *© 2021 Matrix Science* MATRIX SCIENCE

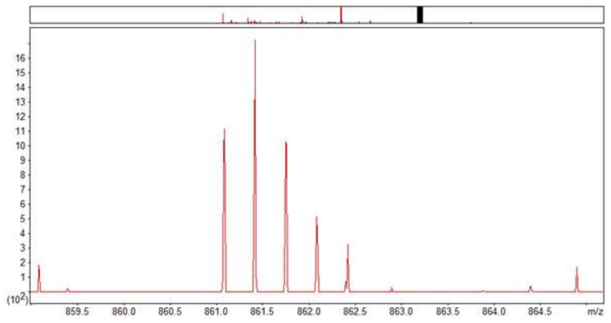
Mascot Distiller 2.8 can use the precursor ion mobility values for identified peptides as a filter on the precursor scans during the XIC peak detection steps of any precursor-based quantitation method, such as SILAC or Label Free. The aim of this is to clean up and improve the quality of the data in the target precursor regions, removing interfering signal, improving accuracy and giving more reliable quantitation results. For a given peptide match, an ion mobility range is calculated from the identified matches and the precursor ion mobility grouping tolerance defined in the processing options. Only MS1 signals from within the calculated ion mobility range are then used during XIC peak detection.

This requires that the data are searched using Mascot Server 2.7 or later, and that the precursor ion mobility values are included in the peak lists.

Unfiltered

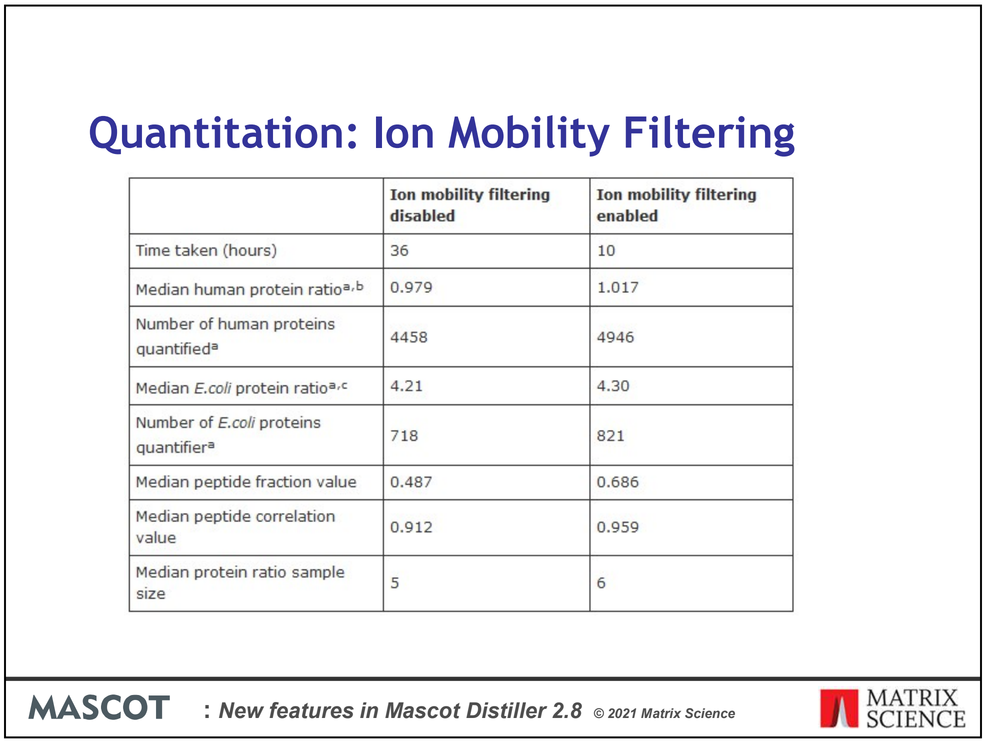


Filtered



Here we can see an unfiltered scan from the precursor region of a peptide. This is a low intensity precursor and, as you can see, there is a lot of noise and interfering signal present.

With filtering enabled, the signal in the precursor area is much cleaner, with very little noise and interference, despite this being a low intensity precursor.



Quantitation: Ion Mobility Filtering

	Ion mobility filtering disabled	**Ion mobility filtering enabled**
Time taken (hours)	36	10
Median human protein ratio[a,b]	0.979	1.017
Number of human proteins quantified[a]	4458	4946
Median *E.coli* protein ratio[a,c]	4.21	4.30
Number of *E.coli* proteins quantifier[a]	718	821
Median peptide fraction value	0.487	0.686
Median peptide correlation value	0.912	0.959
Median protein ratio sample size	5	6

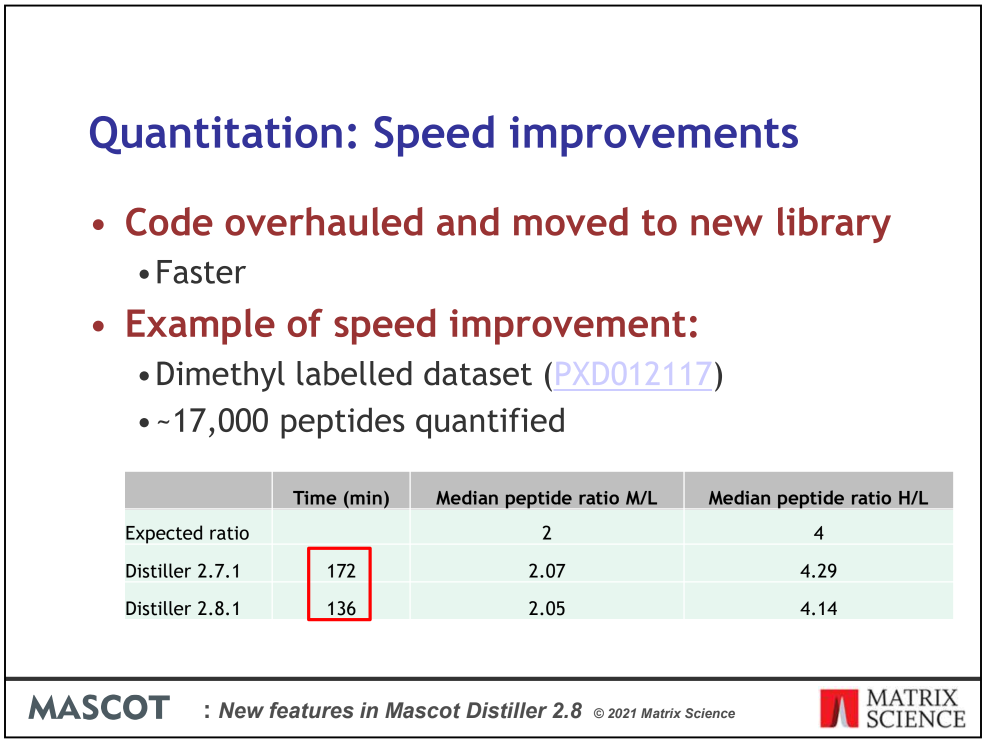
MASCOT : *New features in Mascot Distiller 2.8* © 2021 Matrix Science MATRIX SCIENCE

We took four files, representing two replicates, from the dataset used in the original PASEF paper for timsTOF. These samples consist of a background HeLa sample at a ratio of 1:1 and a spiked in *E.coli* protein digest with a reported median protein ratio of ~4.3. We then carried out quantitation on these samples with and without precursor ion mobility filtering enabled.

As you can see, enabling filtering significantly speeds up quantitation, because less data has to be processed. Median peptide fraction and correlation values also improved, showing the quality and signal to noise of the data are better with filtering enabled, allowing more peptide ratios to pass the quantitation quality thresholds.

The median number of quantitated peptide sequences used to calculate the protein ratios increased from 5 to 6, and more proteins with at least 2 distinct peptide sequences were quantified.

The median protein ratios for both human and *E.coli* proteins are also slightly closer to the expected ratios.



Quantitation: Speed improvements

- Code overhauled and moved to new library
  - Faster
- Example of speed improvement:
  - Dimethyl labelled dataset (PXD012117)
  - ~17,000 peptides quantified

	Time (min)	Median peptide ratio M/L	Median peptide ratio H/L
Expected ratio		2	4
Distiller 2.7.1	172	2.07	4.29
Distiller 2.8.1	136	2.05	4.14

MASCOT : *New features in Mascot Distiller 2.8* © 2021 Matrix Science MATRIX SCIENCE

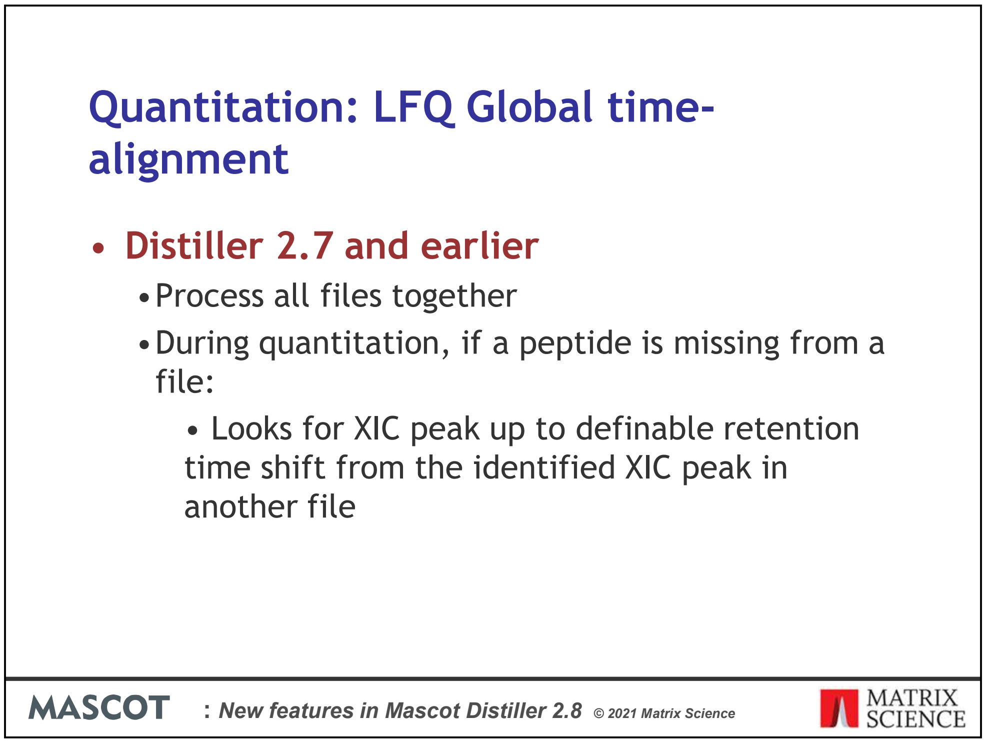
Now we'll take a look at some improvements and new features which benefit all supported instrument data formats.

We've overhauled the quantitation code in Mascot Distiller and moved it into a new library called msquantlib. This has allowed us to tidy up the code, fix some bugs and use a more recent C++ compiler, which means that precursor based quantitation is faster.

As an example of this, we took a dimethyl light, medium and heavy labelled dataset from the PRIDE repository. This is a 'spiked in' dataset, where medium/light has an expected ratio of 2, and heavy/light 4.

We processed the raw files in Distiller, searched using Mascot and then carried out quantitation using Distiller 2.7.1 and 2.8.1 on the same hardware. The dataset contains approximately 17,000 quantifiable peptides. As you can see, the median peptide ratios for Medium/Light and Heavy/Light are similar from both versions of the software, although results are slightly better with Distiller 2.8.

More significantly, Distiller 2.8 was approximately 20% faster, taking 36 minutes less time to complete



Quantitation: LFQ Global time-alignment

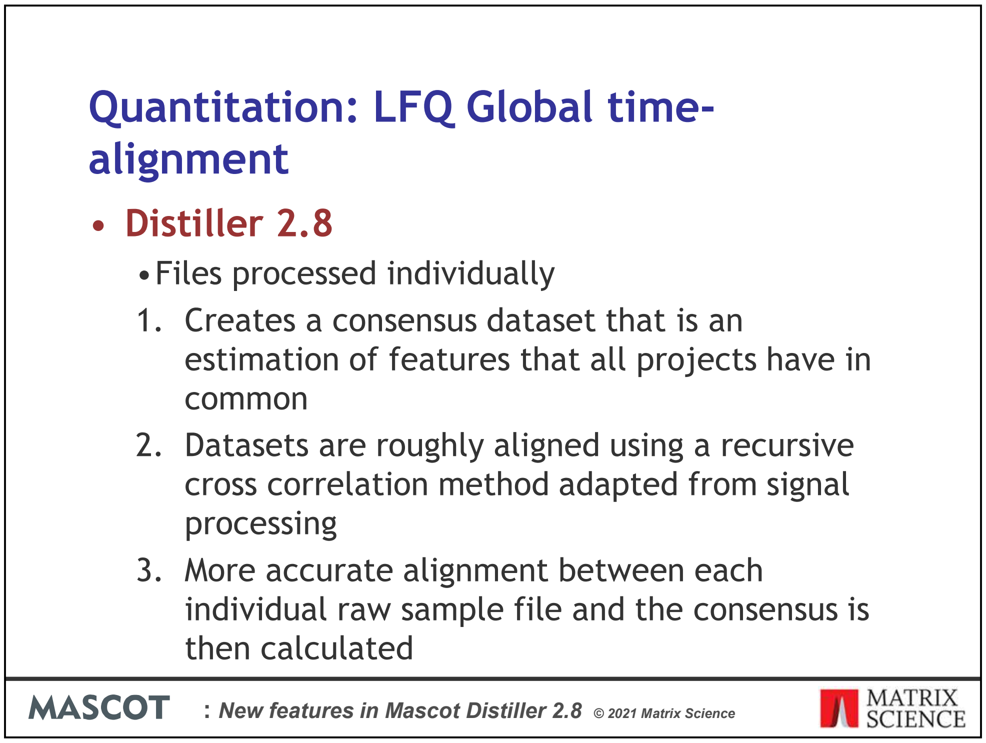
- Distiller 2.7 and earlier
  - Process all files together
  - During quantitation, if a peptide is missing from a file:
    - Looks for XIC peak up to definable retention time shift from the identified XIC peak in another file

MASCOT : *New features in Mascot Distiller 2.8* © 2021 Matrix Science MATRIX SCIENCE

We've implemented a novel global time alignment algorithm for the 'Replicate' label-free quantitation method in Mascot Distiller 2.8. To understand why we've done this, first lets take a look at how replicate quantitation was handled in Mascot Distiller 2.7 and earlier.

To carry out replicate label-free quantitation in Distiller 2.7 or earlier, you had to create a multifile project with all the files processed together (without the 'memory efficient' checkbox checked). The could result in a very large project (depending on the number of samples in the experiment), which could require a very large amount of memory on the workstation to be able to handle it.

During quantitation, if a peptide identified in one or more of the raw sample files is missing from another raw file, Distiller simply searches for an XIC peak in that other file, starting from the identified retention time(s) from other files, up to a user definable limit (500 seconds by default).



Quantitation: LFQ Global time-alignment

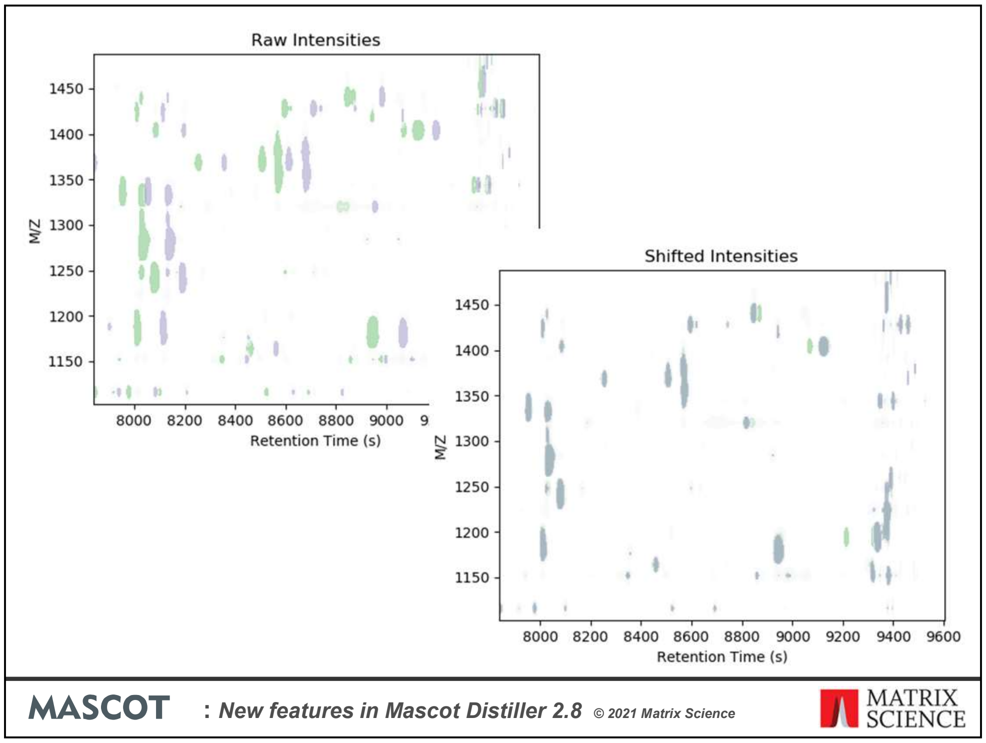
- Distiller 2.8
  - Files processed individually
  1. Creates a consensus dataset that is an estimation of features that all projects have in common
  2. Datasets are roughly aligned using a recursive cross correlation method adapted from signal processing
  3. More accurate alignment between each individual raw sample file and the consensus is then calculated

MASCOT : *New features in Mascot Distiller 2.8* *© 2021 Matrix Science* MATRIX SCIENCE

In Mascot Distiller 2.8, the raw samples can be processed individually, and a 'memory efficient' multifile project created, which reduces the memory overhead. When it carries out replicate quantitation, the first step is then to carry out global time alignment between the files. This is then used to locate the regions to look for XICs for any identified peptides missing from a particular file.

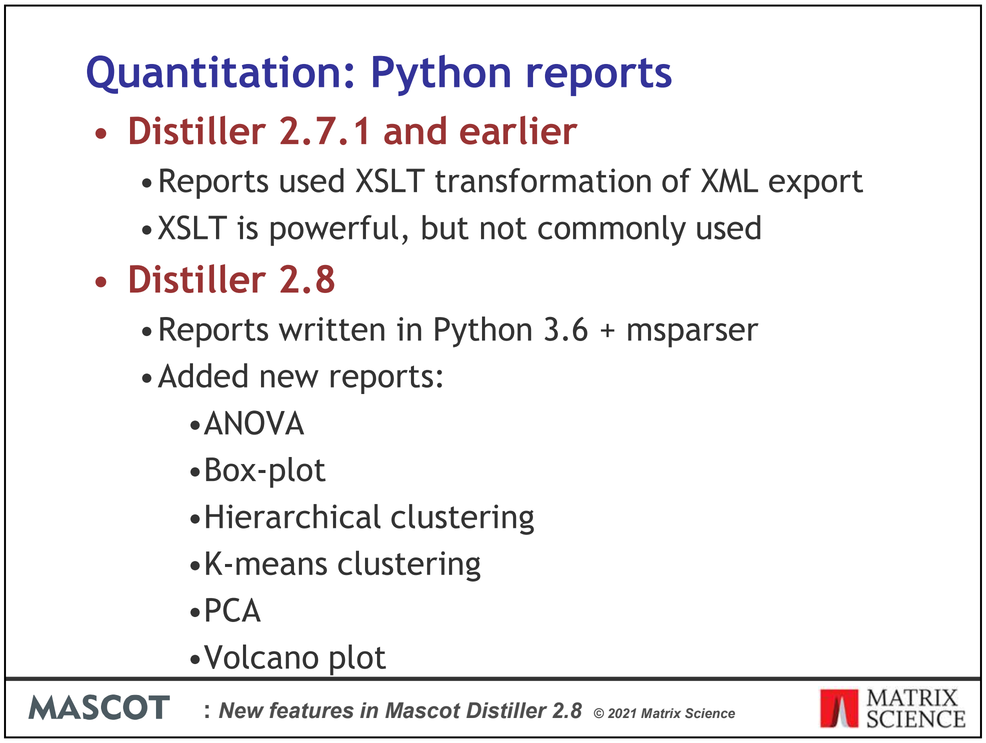
Briefly, the approach used is as follows:

1. A 'consensus' dataset is created which is an estimation of features that all projects have in common
2. The individual datasets are then roughly aligned to the consensus
3. A more accurate alignment between each individual raw sample file and the consensus is then calculated



There is a separate presentation looking in more detail at this, but to show the effect of the alignment, here we have MS1 features from two files in a label-free project plotted, with retention time on the X-axis and m/z on the Y-axis. Before alignment, we can see a clear shift in feature retention times between the two files. After time alignment, the common features between the two files are now clearly overlapping.

Note that this plot was created outside of Distiller for illustration, and is not currently available as a report.



Quantitation: Python reports

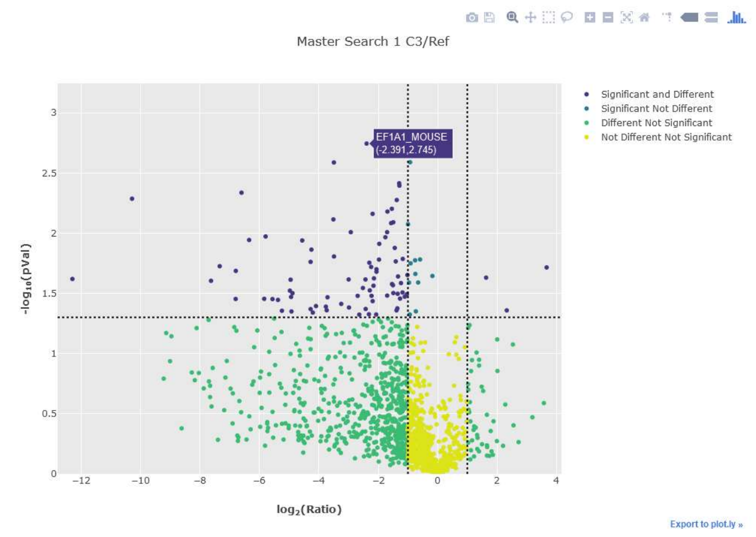
- Distiller 2.7.1 and earlier
  - Reports used XSLT transformation of XML export
  - XSLT is powerful, but not commonly used
- Distiller 2.8
  - Reports written in Python 3.6 + msparser
  - Added new reports:
    - ANOVA
    - Box-plot
    - Hierarchical clustering
    - K-means clustering
    - PCA
    - Volcano plot

MASCOT : *New features in Mascot Distiller 2.8* *© 2021 Matrix Science* MATRIX SCIENCE

In Mascot Distiller 2.7 and earlier, reporting of quantitation results was handled using XSLT transformation of the XML export of the quantitation results. XSLT is a powerful language, but it’s not commonly used and you’d probably struggle to find someone in the lab who was familiar enough with it to write custom reports.

In Mascot Distiller 2.8 reports are written in Python 3.6 and use msparser to access the search and quantitation results. Python is a commonly used, powerful programming language and using msparser provides simple access to the search and quantitation results.

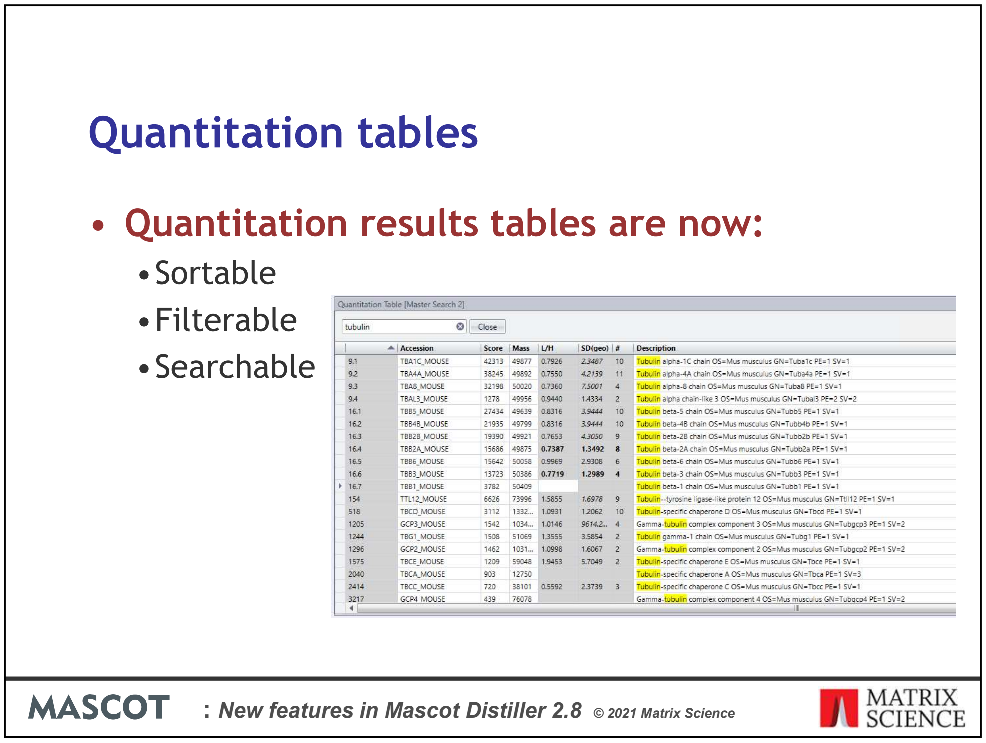
We’ve taken the opportunity to add some new reports to Distiller, so you’ll now find options to generate ANOVA, Box-Plot, various clustering reports including PCA, and a volcano plot.



Accession	Log2(RatioValue)	-Log10(pVal)	Description
EF1A1_MOUSE	-2.391	2.745	Elongation factor 1-alpha 1 OS=Mus musculus OX=10090 GN=Eef1a1 PE=1 SV=3
FABP5_MOUSE	-3.494	2.588	Fatty acid-binding protein 5 OS=Mus musculus OX=10090 GN=Fabp5 PE=1 SV=3
LDHA_MOUSE	-1.295	2.415	L-lactate dehydrogenase A chain OS=Mus musculus OX=10090 GN=Ldha PE=1 SV=3
RL28_MOUSE	-1.287	2.395	60S ribosomal protein L28 OS=Mus musculus OX=10090 GN=Rpl28 PE=1 SV=2

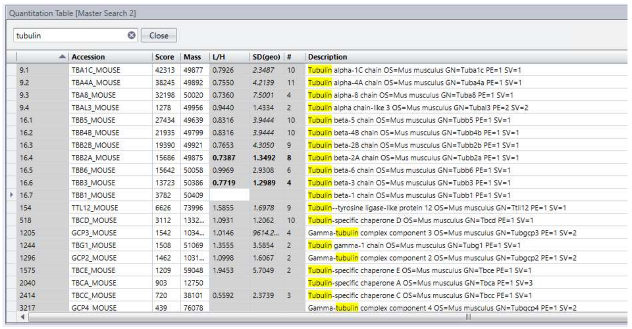
MASCOT : *New features in Mascot Distiller 2.8* *© 2021 Matrix Science* MATRIX SCIENCE

Here's an example of the Volcano plot report for one ratio of a label-free dataset. We've exported using the 'interactive javascript' option so that hovering over a point on the graph displays the associated protein accession. This information is also presented in a set of tables below the graph.



Quantitation tables

- Quantitation results tables are now:
  - Sortable
  - Filterable
  - Searchable



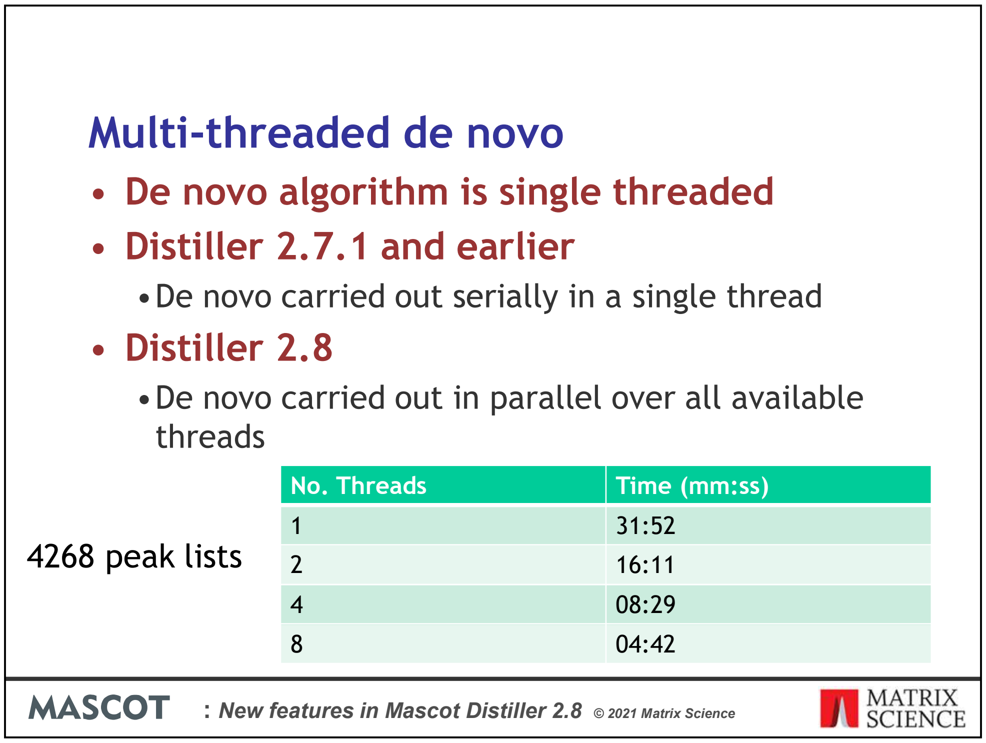
Quantitation Table [Master Search 2]

tubulin

	Accession	Score	Mass	L/H	SD(geo)	#	Description
9.1	TBA1C_MOUSE	42313	49877	0.7926	*2.3487*	10	Tubulin alpha-1C chain OS=Mus musculus GN=Tuba1c PE=1 SV=1
9.2	TBA4A_MOUSE	38245	49892	0.7550	*4.2139*	11	Tubulin alpha-4A chain OS=Mus musculus GN=Tuba4a PE=1 SV=1
9.3	TBA8_MOUSE	32198	50020	0.7360	*7.5001*	4	Tubulin alpha-8 chain OS=Mus musculus GN=Tuba8 PE=1 SV=1
9.4	TBAL3_MOUSE	1278	49956	0.9440	1.4334	2	Tubulin alpha chain-like 3 OS=Mus musculus GN=Tubal3 PE=2 SV=2
16.1	TBB5_MOUSE	27434	49639	0.8316	*3.9444*	10	Tubulin beta-5 chain OS=Mus musculus GN=Tubb5 PE=1 SV=1
16.2	TBB4B_MOUSE	21935	49799	0.8316	*3.9444*	10	Tubulin beta-4B chain OS=Mus musculus GN=Tubb4b PE=1 SV=1
16.3	TBB2B_MOUSE	19390	49921	0.7653	*4.3050*	9	Tubulin beta-2B chain OS=Mus musculus GN=Tubb2b PE=1 SV=1
16.4	TBB2A_MOUSE	15686	49875	**0.7387**	**1.3492**	**8**	Tubulin beta-2A chain OS=Mus musculus GN=Tubb2a PE=1 SV=1
16.5	TBB6_MOUSE	15642	50058	0.9969	2.9308	6	Tubulin beta-6 chain OS=Mus musculus GN=Tubb6 PE=1 SV=1
16.6	TBB3_MOUSE	13723	50386	**0.7719**	**1.2989**	**4**	Tubulin beta-3 chain OS=Mus musculus GN=Tubb3 PE=1 SV=1
16.7	TBB1_MOUSE	3782	50409				Tubulin beta-1 chain OS=Mus musculus GN=Tubb1 PE=1 SV=1
154	TTL12_MOUSE	6626	73996	1.5855	*1.6978*	9	Tubulin--tyrosine ligase-like protein 12 OS=Mus musculus GN=Ttll12 PE=1 SV=1
518	TBCD_MOUSE	3112	1332...	1.0931	1.2062	10	Tubulin-specific chaperone D OS=Mus musculus GN=Tbcd PE=1 SV=1
1205	GCP3_MOUSE	1542	1034...	1.0146	*9614.2...*	4	Gamma-tubulin complex component 3 OS=Mus musculus GN=Tubgcp3 PE=1 SV=2
1244	TBG1_MOUSE	1508	51069	1.3555	3.5854	2	Tubulin gamma-1 chain OS=Mus musculus GN=Tubg1 PE=1 SV=1
1296	GCP2_MOUSE	1462	1031...	1.0998	1.6067	2	Gamma-tubulin complex component 2 OS=Mus musculus GN=Tubgcp2 PE=1 SV=2
1575	TBCE_MOUSE	1209	59048	1.9453	5.7049	2	Tubulin-specific chaperone E OS=Mus musculus GN=Tbce PE=1 SV=1
2040	TBCA_MOUSE	903	12750				Tubulin-specific chaperone A OS=Mus musculus GN=Tbca PE=1 SV=3
2414	TBCC_MOUSE	720	38101	0.5592	2.3739	3	Tubulin-specific chaperone C OS=Mus musculus GN=Tbcc PE=1 SV=1
3217	GCP4_MOUSE	439	76078				Gamma-tubulin complex component 4 OS=Mus musculus GN=Tubgcp4 PE=1 SV=2

There are some new features available on the protein and peptide quantitation tables. Tables are now sortable – either by clicking on the column header or from the column context menu - and they are also searchable and filterable.

Here, I've enabled the table's search panel and searched for tubulin on the protein quantitation results table. The table has been filtered so that only the matching protein rows are displayed, with the matching text highlighted in yellow.



Multi-threaded de novo

- De novo algorithm is single threaded
- Distiller 2.7.1 and earlier
    - De novo carried out serially in a single thread
- Distiller 2.8
    - De novo carried out in parallel over all available threads

4268 peak lists

No. Threads	Time (mm:ss)
1	31:52
2	16:11
4	08:29
8	04:42

MASCOT : *New features in Mascot Distiller 2.8* © 2021 Matrix Science MATRIX SCIENCE

And finally, we've made the de novo search feature in Distiller multi-threaded

The de novo algorithm itself is single threaded – that is a single thread is used to process a peaklist.

In Distiller 2.7.1 and earlier, if you requested de novo on a range of peaklists, de novo was carried out across those peaklists serially in a single thread, regardless of how many CPU cores you had available on the workstation.

In Distiller 2.8, de novo is now carried out in parallel using all available threads. This can significantly improve performance if you are carrying out de novo on a large number of peaklists.

As an example of this, here we have the timings for carrying out de novo on 4268 peak lists from a dataset using 1, 2, 4 or 8 threads to carry out de novo on a system with 8 'real' cores available. As you can see, each time we double the number of threads, we almost, but not quite, halve the processing time.