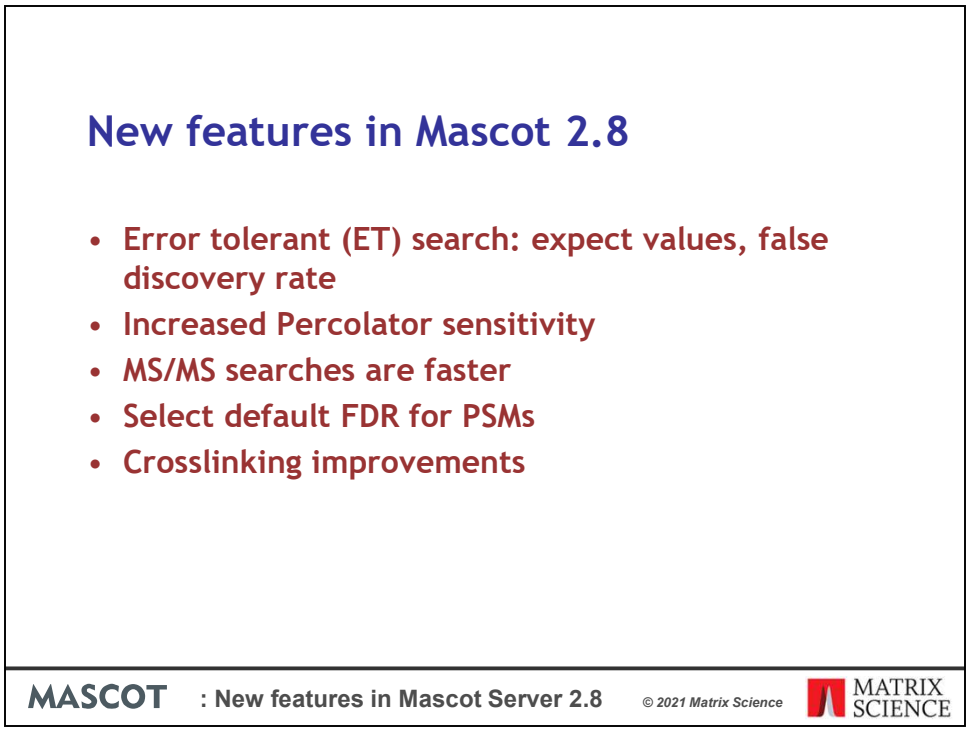
New features in Mascot Server 2.8

Ville Koskinen



Mascot Server 2.8 was released in July 2021. Here's a summary of the new features and significant changes.

Error tolerant searching has long been part of Mascot. We've added a new statistical model for ET expect values, and you can also submit a target-decoy ET search to estimate false discovery rate.

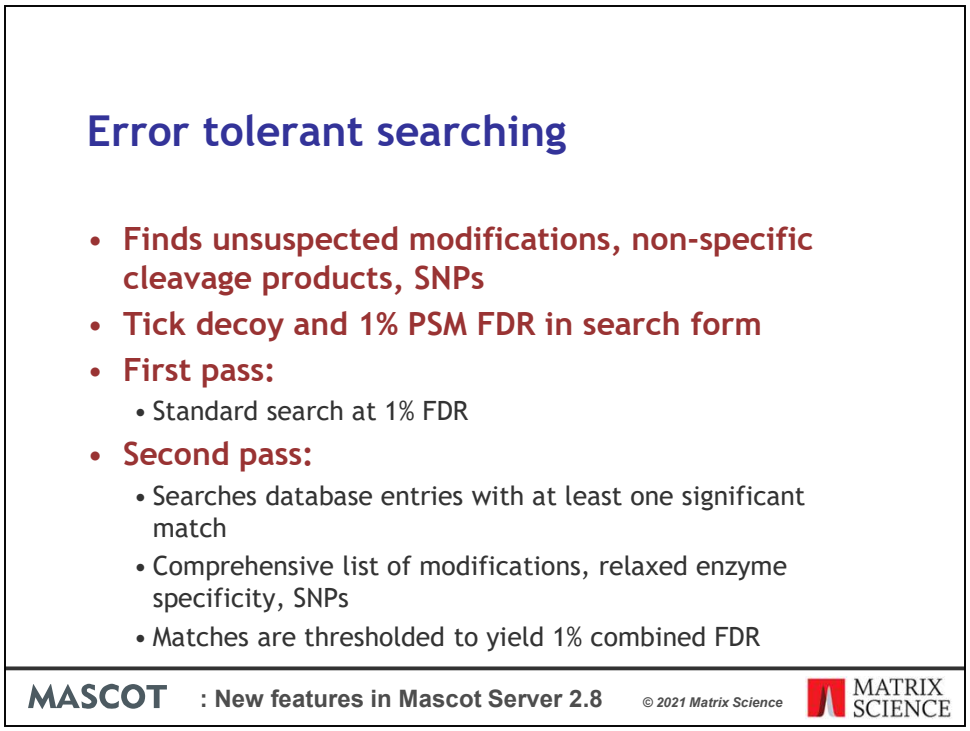
We've added new computed peptide features for Percolator, which increases Percolator sensitivity in most data sets. It's especially beneficial for endogenous peptides.

MS/MS database searches in the new version are about 20-30% faster with medium to large searches. This is due to removing disk access bottlenecks.

You can now choose a default false discovery rate for peptide-spectrum matches when you submit a search.

We've added a configuration editor for crosslinking methods as well as exporting crosslinked search results in CSV and XML format. The speed and memory usage of

crosslinked searches has also been improved.



The Mascot error tolerant search is the most efficient way to find unsuspected modifications, non-specific cleavage products and sequence variants.

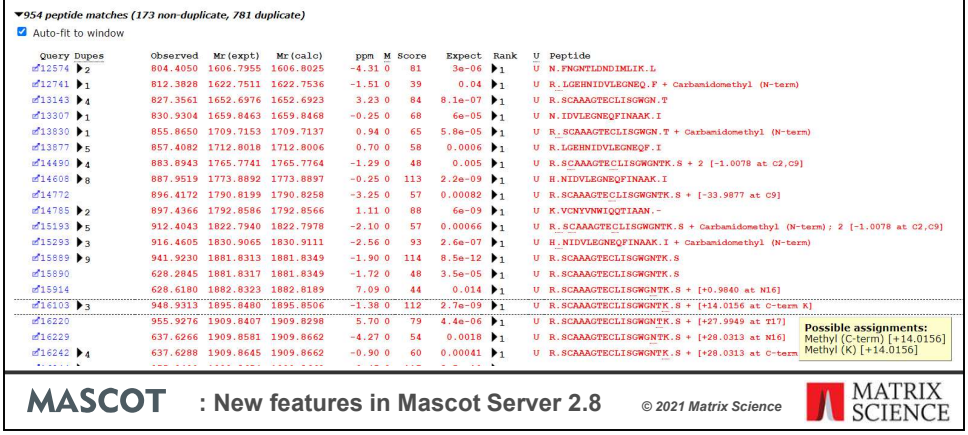
Error tolerant searching is not new in itself. In version 2.8, we've reviewed the statistical model. You can now submit the ET search as a target-decoy search and select the desired false discovery rate in the search form. This gives a solid, empirical basis for the statistics.

The first pass search is a standard database search, and the matches are thresholded to the selected FDR, by default 1% FDR.

In the second pass, Mascot only searches database entries which have at least one significant match from the first pass at the chosen FDR level. Then, as in previous versions, Mascot iterates through a comprehensive list of modifications, uses relaxed enzyme specificity and tries single residue substitutions or single nucleotide substitutions, insertions and deletions. Finally, the first and second pass results are merged and the second pass matches are thresholded to yield 1% combined FDR.

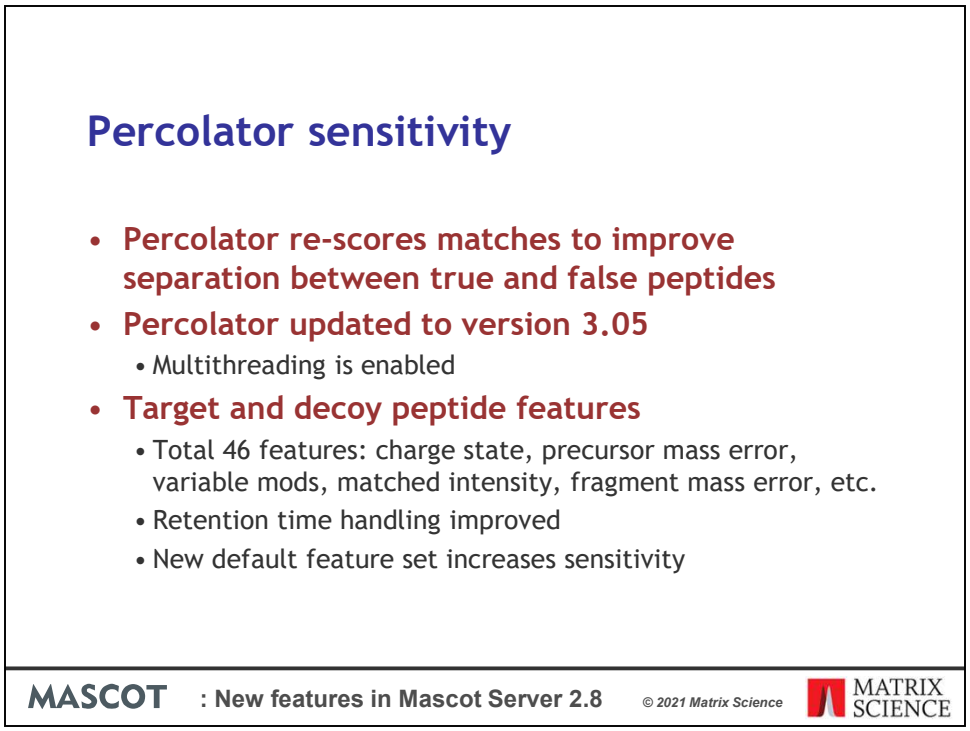
Error tolerant searching

- **ASMS 2021 presentation: "Statistical Significance in Error Tolerant Search Results"**



Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
12574	2	804.4050	1606.7955	1606.8025	-4.31	0	81	3e-06	1	U	N.FNGNTLDNDIMLIK.L
12741	1	812.3828	1622.7511	1622.7536	-1.51	0	39	0.04	1	U	R.LGEHNIDVLEGNEQ.F + Carbamidomethyl (N-term)
13143	4	827.3561	1652.6976	1652.6923	3.23	0	84	8.1e-07	1	U	R.SCAAAGTECLISGWGN.T
13307	1	830.9304	1659.8463	1659.8468	-0.25	0	68	6e-05	1	U	N.IDVLEGNEQFINAAK.I
13830	1	855.8650	1709.7153	1709.7137	0.94	0	65	5.8e-05	1	U	R.SCAAAGTECLISGWGN.T + Carbamidomethyl (N-term)
13877	5	857.4082	1712.8018	1712.8006	0.70	0	58	0.0006	1	U	R.LGEHNIDVLEGNEQF.I
14490	4	883.8943	1765.7741	1765.7764	-1.29	0	48	0.005	1	U	R.SCAAAGTECLISGWGNTK.S + 2 [-1.0078 at C2,C9]
14608	8	887.9519	1773.8892	1773.8897	-0.25	0	113	2.2e-09	1	U	H.NIDVLEGNEQFINAAK.I
14772		896.4172	1790.8199	1790.8258	-3.25	0	57	0.00082	1	U	R.SCAAAGTECLISGWGNTK.S + [-33.9877 at C9]
14785	2	897.4366	1792.8586	1792.8566	1.11	0	88	6e-09	1	U	K.VCNYVNWIQQTIAAN.-
15193	5	912.4043	1822.7940	1822.7978	-2.10	0	57	0.00066	1	U	R.SCAAAGTECLISGWGNTK.S + Carbamidomethyl (N-term); 2 [-1.0078 at C2,C9]
15293	3	916.4605	1830.9065	1830.9111	-2.56	0	93	2.6e-07	1	U	H.NIDVLEGNEQFINAAK.I + Carbamidomethyl (N-term)
15889	9	941.9230	1881.8313	1881.8349	-1.90	0	114	8.5e-12	1	U	R.SCAAAGTECLISGWGNTK.S
15890		628.2845	1881.8317	1881.8349	-1.72	0	48	3.5e-05	1	U	R.SCAAAGTECLISGWGNTK.S
15914		628.6180	1882.8323	1882.8189	7.09	0	44	0.014	1	U	R.SCAAAGTECLISGWGNTK.S + [+0.9840 at N16]
16103	3	948.9313	1895.8480	1895.8506	-1.38	0	112	2.7e-09	1	U	R.SCAAAGTECLISGWGNTK.S + [+14.0156 at C-term K]
16220		955.9276	1909.8407	1909.8298	5.70	0	79	4.4e-06	1	U	R.SCAAAGTECLISGWGNTK.S + [+27.9949 at T17]
16229		637.6266	1909.8581	1909.8662	-4.27	0	54	0.0018	1	U	R.SCAAAGTECLISGWGNTK.S + [+28.0313 at N16]
16242	4	637.6288	1909.8645	1909.8662	-0.90	0	60	0.00041	1	U	R.SCAAAGTECLISGWGNTK.S + [+28.0313 at C-term

Here is a screenshot of search results of an error tolerant search. The highlighted row is a high-scoring ET match with C-terminal methylation, and the Expect column displays the statistical confidence. For more details, have a look at our ASMS 2021 presentation "Statistical Significance in Error Tolerant Search Results".



The next improvement is also in target-decoy searching. Mascot ships with Percolator, which is a semi-supervised machine learning tool. Percolator re-scores database matches to improve separation between true and false matches.

Percolator has been shipped with Mascot for many years. In version 2.8, we've updated the Percolator executable to the latest version, 3.05. The Percolator training phase is now multithreaded, so will use more than one CPU core.

Percolator training depends on computed features extracted from target and decoy matches. We've added new features in Mascot 2.8, which supports 46 features in total. This includes features like charge state, precursor mass error, variable modifications, matched intensity, fragment mass error and so on. We've also improved the handling of retention time as a feature.

Mascot ships with a new default feature set. The feature set was designed so that you get an improvement in sensitivity in most data sets.

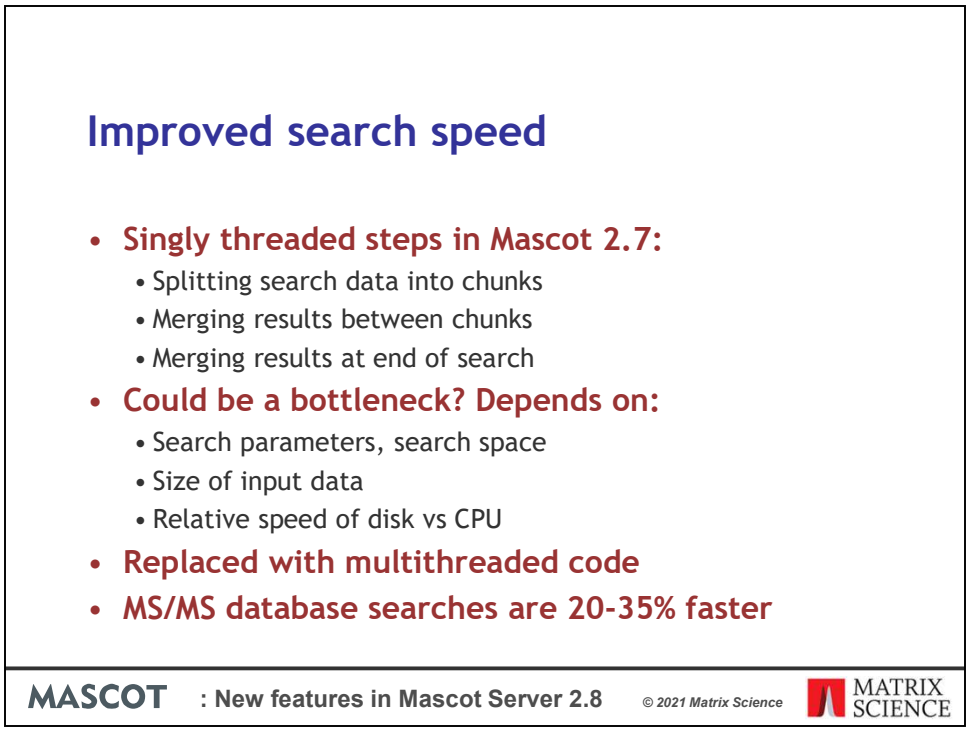
Percolator sensitivity

- **September 2021 blog: "Identify more HLA peptides"**
  - 60% more PSMs, 30% more sequences than Mascot 2.7

	Target PSMs	PSM FDR	Sequences	Sequence FDR
Mascot 2.7	8669	0.99%	1105	2.71%
+ Percolator	22403	1.02%	1929	4.67%
+ RT enabled	22602	0.92%	1928	4.25%
Mascot 2.8	8669	0.99%	1105	2.71%
+ Percolator	31744	1.00%	2349	4.21%
+ RT enabled	36338	1.00%	2496	4.53%

MASCOT : New features in Mascot Server 2.8 MATRIX SCIENCE

The improvement in sensitivity is particularly good with endogenous peptides. Our September 2021 blog article uses a library of HLA peptides to illustrate the improvement. In this data set, the new feature set gives 60% more PSMs and 30% more peptide sequences at the same FDR compared to Mascot 2.7. Enabling the retention time feature in the previous version made little difference in this data set, whereas it gives an additional boost in Mascot 2.8.



Database searches in Mascot 2.8 are a bit faster than in the previous version. Mascot used to have a few steps where only a single processing thread is active, no matter the size of your licence. These were: splitting search data into chunks at the beginning of the search; merging intermediate results into a temporary file between chunks; and merging results at the end of the search.

Whether the steps are actually a bottleneck depends on the search parameters, search space and size of the input data. The relative speed of the disk compared to the CPU is also a factor. As a rough guide, if the search has tight mass tolerances and few variable modifications, a reasonable amount of time could be spent on disk operations. When the search is disk bound, the singly threaded steps can be a bottleneck.

Conversely, an error tolerant search is likely to be CPU bound and relatively little time is spent on disk operations. In this case, the singly threaded steps are not a bottleneck.

We replaced the singly threaded steps with multithreaded code and benchmarked a range of typical data sets. The change is an overall improvement: disk bound MS/MS database searches are now 20 to 35 percent faster. Whether a specific search is faster in the new version depends on the factors mentioned earlier.

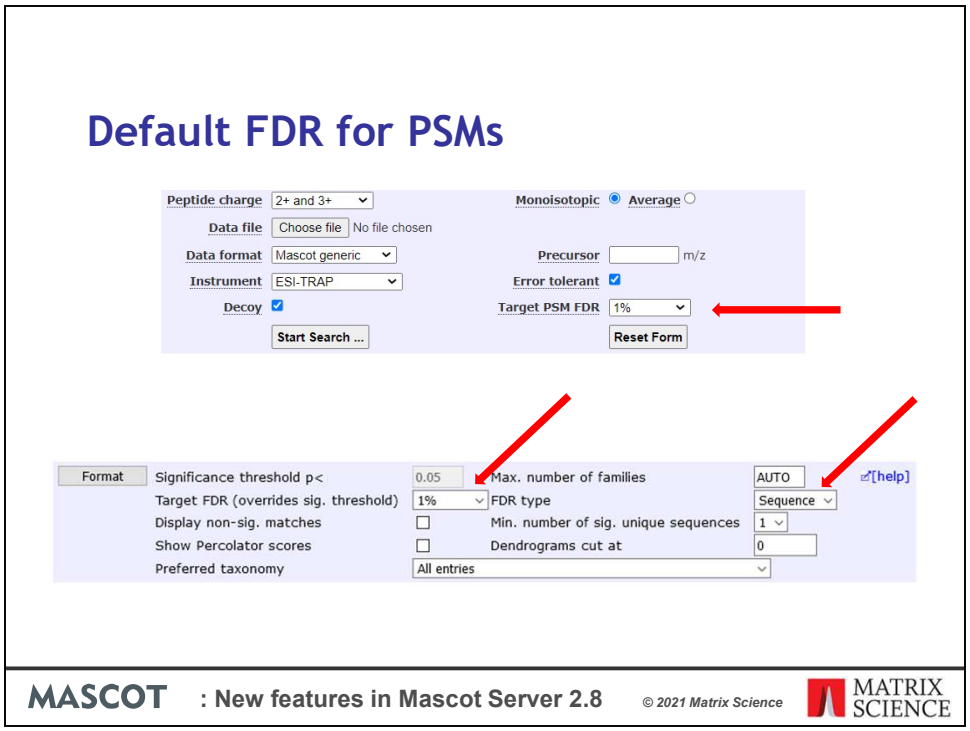
Improved search speed

- **August 2021 blog article**
  - 683,905 spectra, human proteome
  - Typical 4-core Intel Core i7

Mascot version	Disk type	Search time/seconds	Relative to 2.7
2.7	HDD	1435	100%
2.8	HDD	1050	73%
2.7	SSD	1137	100%
2.8	SSD	877	77%

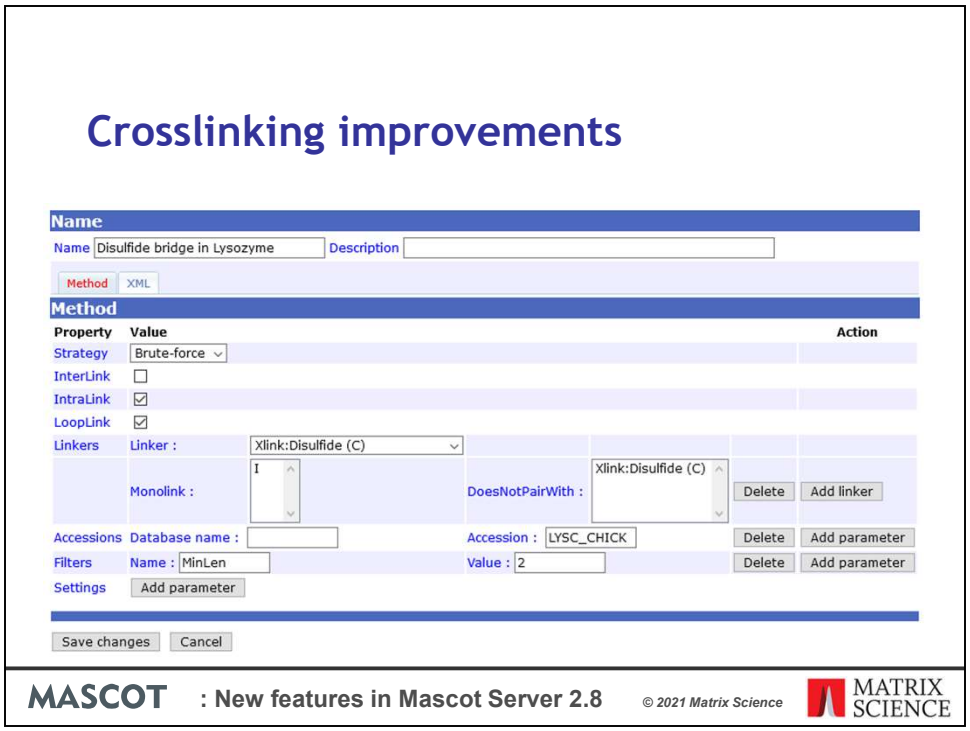
Our August 2021 blog article has more detail about the change and describes one of the benchmarking data sets. This is a search of the human proteome of 683,905 spectra. The search was run on a typical 4-core Intel Core i7 system using a 1-CPU Mascot licence. The search had one variable modification and tight mass tolerances.

As you can see in the table, the new version is faster by a good margin on both traditional hard disks and solid state drives. The improvement also applies multi-CPU licences, cluster installations and systems with RAID arrays.



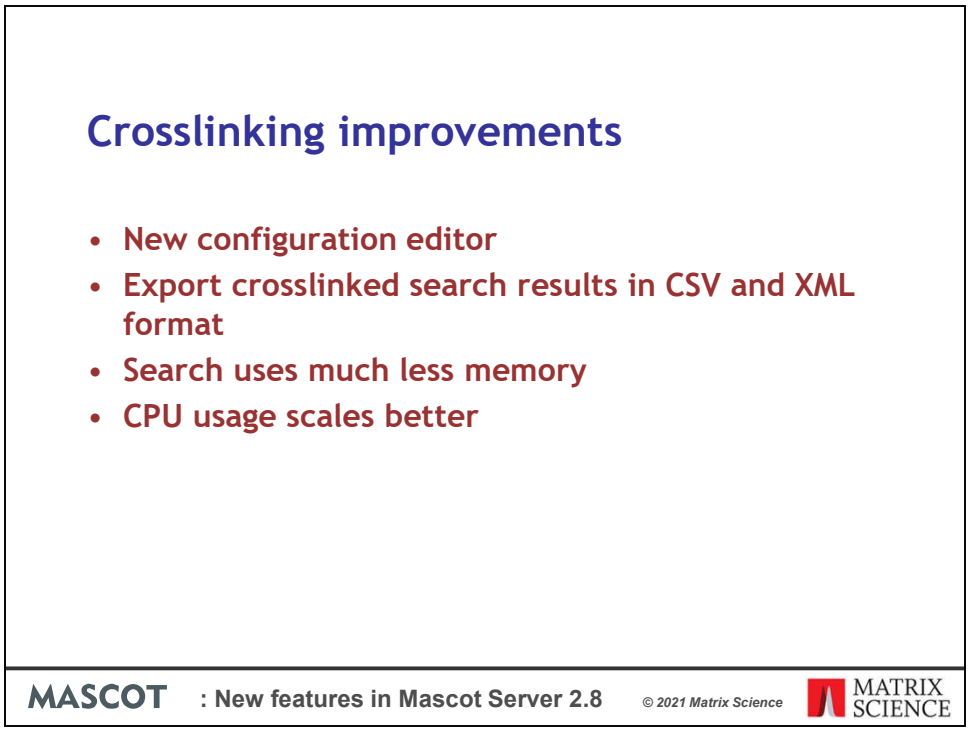
Mascot 2.8 has a new search form control for target-decoy searches. You can select a target FDR for peptide-spectrum matches from a dropdown menu. The setting applies to both standard searches and error tolerant searches. In standard searches, results are automatically thresholded to the selected FDR when the report loads. In error tolerant searches, Mascot additionally thresholds first-pass matches to the selected FDR before running the second pass search.

Protein Family Summary has two new format controls. The target FDR dropdown menu allows selecting a different FDR. The FDR type dropdown allows choosing either PSM FDR or sequence FDR.



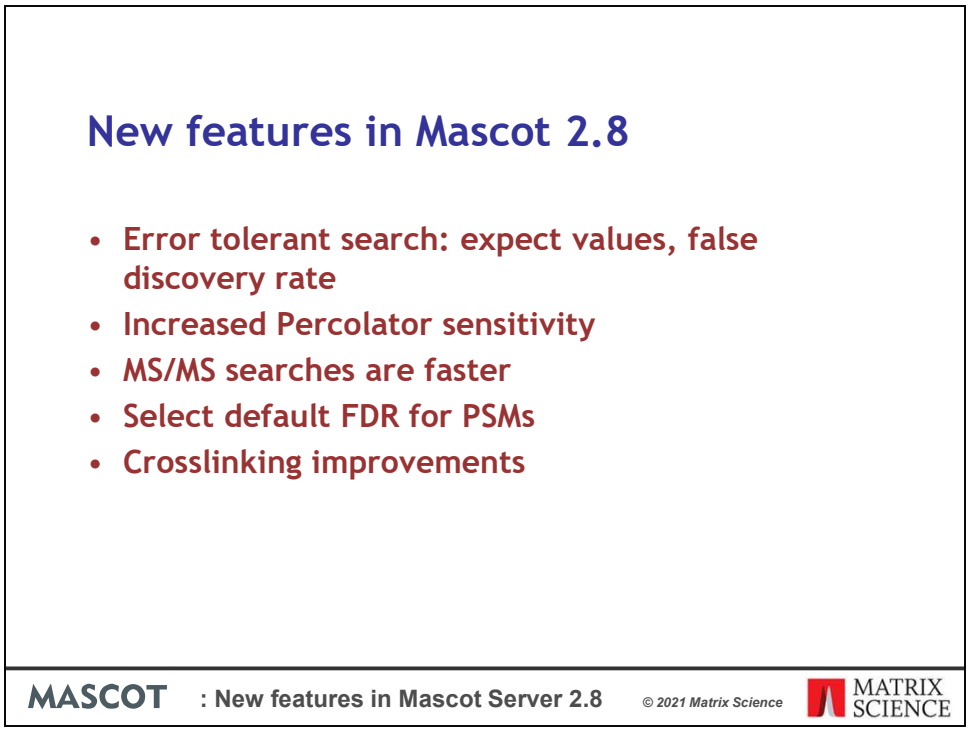
Searching intact crosslinked peptides was added in Mascot 2.7.

In Mascot 2.8, we've made a few more improvements. The first is a new configuration editor for crosslinking methods. The user interface is similar to the quantitation method editor. You can tick the boxes to enable or disable interlinking, intralinking and looplinking. The linkers are available in a dropdown menu, and any monolinks or directionality constraints can be chosen in the multiselect boxes. At the bottom are the accession settings and other filters.



You can now export integrated search results in CSV and XML format. The output file contains both linear and crosslinked matches, including any monolinked peptides.

The crosslinked part of the database search uses much less memory than Mascot 2.7. We've also improved CPU scaling with crosslinked searches. If your search uses only intralinking, there should now be no limit on the number of proteins that can be intralinked. The crosslinking method still specifies a soft limit called MaxProteins, which is a safety valve for interlinked searches.



Thank you for your attention.