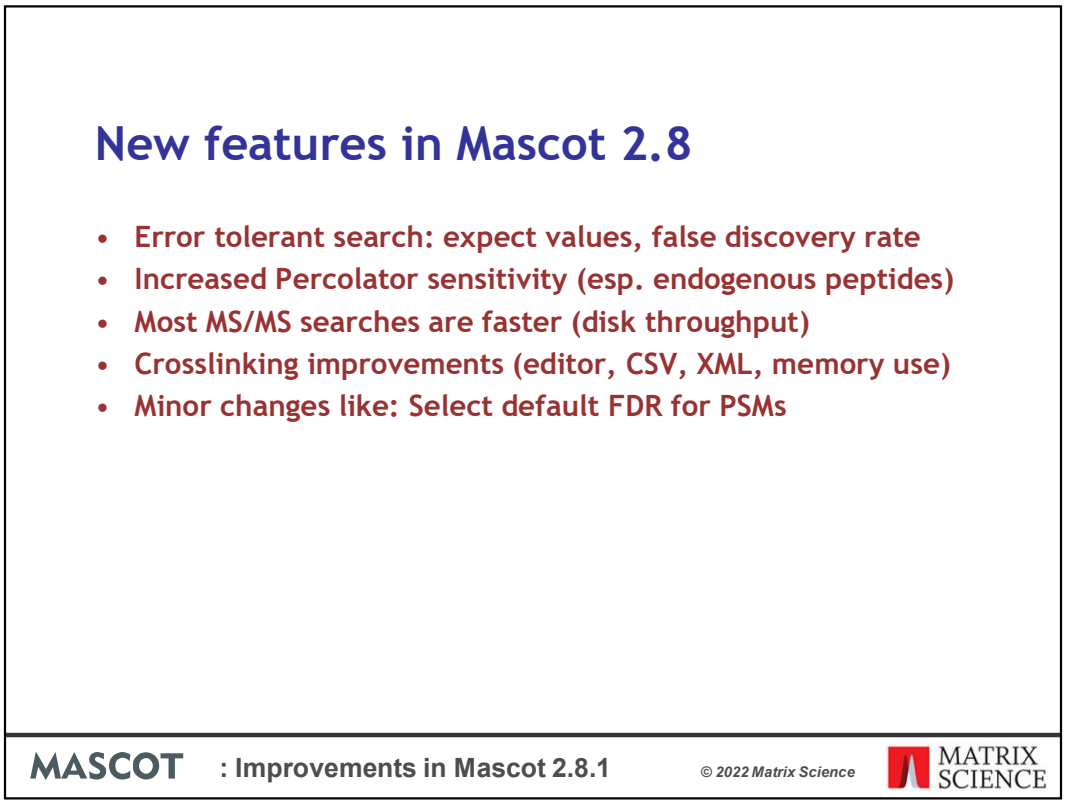
NCBIprot, mzIdentML 1.2 and other improvements in Mascot Server 2.8.1

Ville Koskinen



Mascot Server 2.8 was released in July 2021. We presented the new features in an online presentation at ASMS 2021, so I'll only give the highlights on this slide.

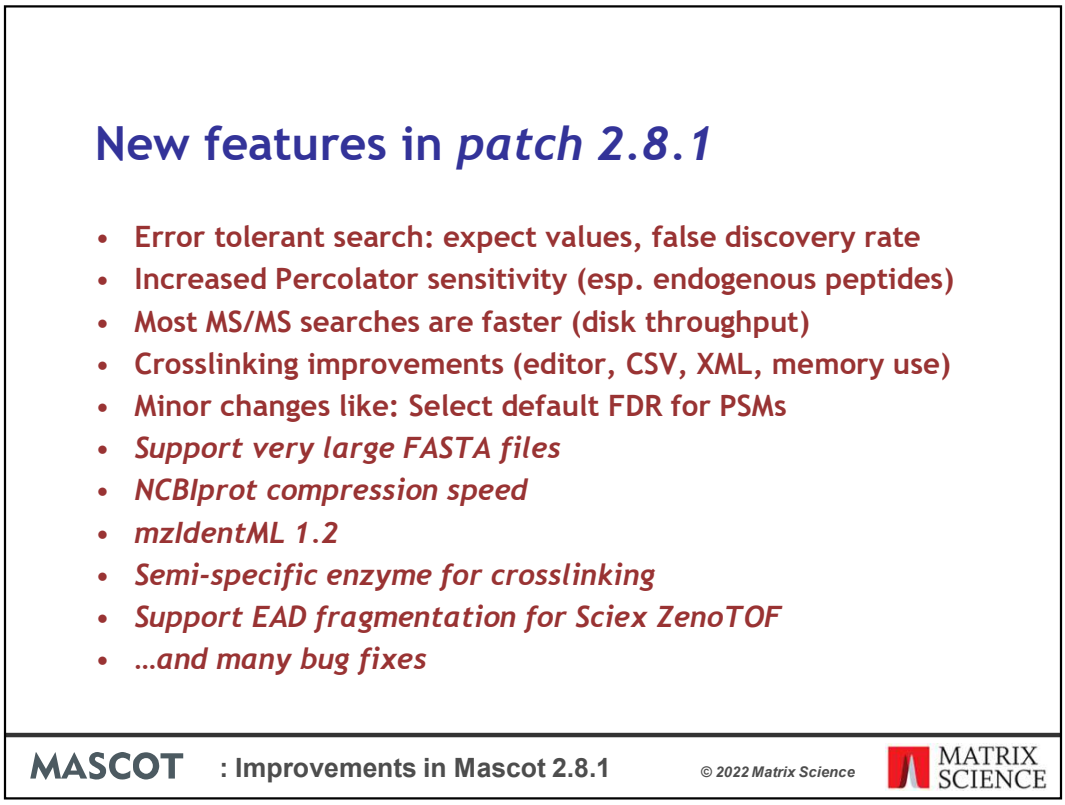
Error tolerant searching has long been part of Mascot. We've added a new statistical model for ET expect values, and you can also submit a target-decoy ET search to estimate false discovery rate.

We 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.

We 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 was also improved.

There were also several minor changes, such as: You can now choose a default false discovery rate for peptide-spectrum matches when you submit a search.



The latest version is 2.8.1, which is a patch released in March 2022. It includes a number of bug fixes and improvements.

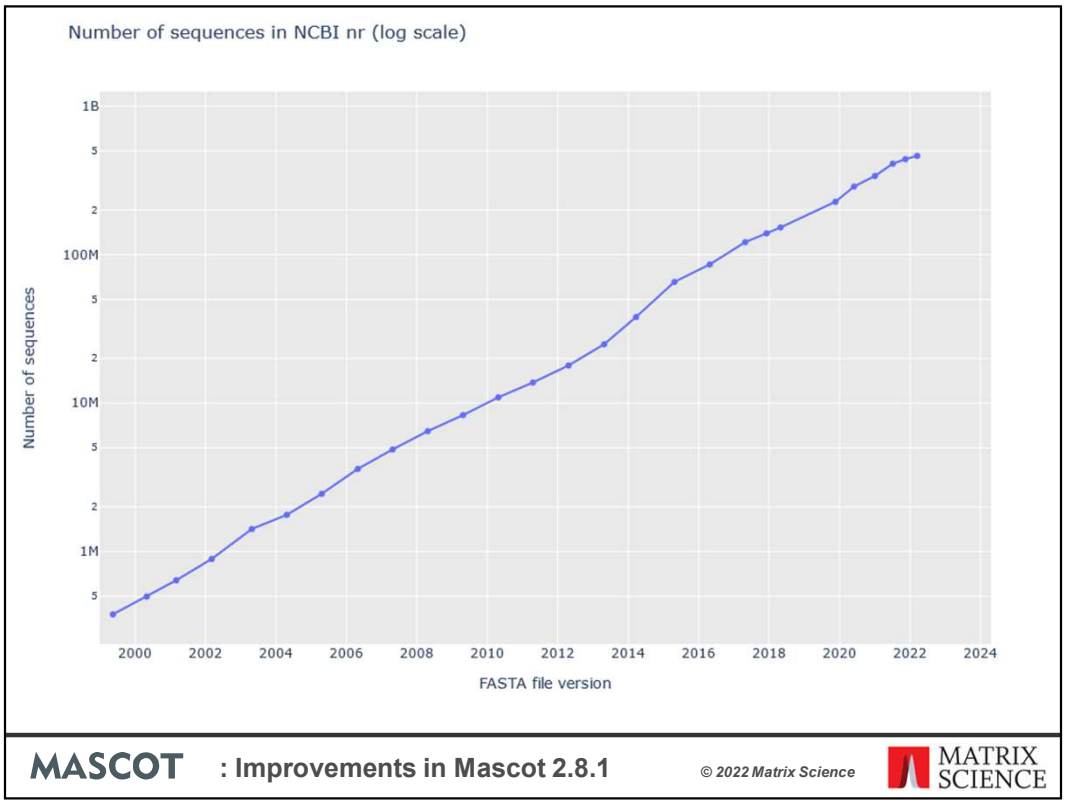
We've improved the user experience with NCBIprot. This is a very large database that doubles in size every couple of years, and earlier versions of Mascot really struggled with it.

Search results can now be exported in mzIdentML 1.2 format. Mascot has long had support for mzIdentML 1.1. New in the 1.2 standard is support for crosslinked search results.

We've also made it possible to submit a crosslink search using a semi-specific enzyme like semi-trypsin.

The patch adds an instrument definition for electron-assisted dissociation or EAD fragmentation. This is the type of fragmentation from the new Sciex ZenoTOF instruments.

And there are many smaller bug fixes.



NCBI nr is the largest public protein sequence database. It's available in Mascot as the NCBIprot predefined definition. I've plotted the size of the database as a time series, where the y axis uses log scale and shows the number of protein sequences. The x axis spans from 1999 to the present day.

Oddly, there isn't an official graph like this available, so we've constructed the time series from data on our public website. The free to use Mascot service started in 1999, and it has had nr as an available database since then. When a search is submitted, the FASTA file version and number of sequences are saved in the results file header. That's where the counts come from.

The growth is linear on log scale so actually nr is roughly doubling in size every 2 years. The slope is almost constant the past 20 years, so it should reach 1 billion protein sequences in 2024.

Nr is compiled from multiple databases. The main growth seems to come from NCBI RefSeq, which is a comprehensive collection of genomic DNA, transcripts and protein sequences.

Very large FASTA files

DB	Date	Sequences	Accessions	2.8.0	2.8.1
NCBIprot	January 2021	339M	605M	Yes	Yes
	July 2021	410M	730M	No	Yes
	November 2021	440M	762M	No	Yes
	March 2022	463M	796M	No	Yes
TrEMBL	March 2022	230M	230M	Yes	Yes
UniRef100	March 2022	300M	300M	Yes	Yes

- 2.8.0: 32-bit limit when taxonomy defined
- 2.8.1: unlimited FASTA size when taxonomy defined

Last year, we discovered that once the database reached a critical size, Mascot could no longer bring it online. The breakpoint is around 370 million sequences. Mascot 2.8 works fine when the size is 339 million sequences and fails when it's 410 million.

The fix was quite small but it took long to find. When Mascot brings a database online, it creates several compressed index files. One of these index files lists the taxonomy IDs for each protein sequence. Nr is unusual in that a sequence can have multiple protein accessions, so each sequence has multiple taxonomy IDs. The fault was, Mascot used a 32-bit integer when reading the index file. It was OK up to about 650 million taxonomy IDs but no more.

We've fixed the fault in the patch release. As you can see from the table, the new version has no limit on database size.

The other large databases predefined in Mascot are TrEMBL and UniRef100. These are relatively smaller than NCBIprot and they don't grow as quickly. Both work in the earlier version and will continue to work until they reach the critical size in a few years.

NCBIprot compression speed

Type	CPU	RAM	Disk	Platform	2.8.0, 339M sequences	2.8.01, 440M sequences
Server	Xeon 3.4GHz	128GB	RAID10, 10k rpm	Linux	5h	6h
Workstation	Core i7 3.2GHz	64GB	SSD + HDD	Windows	240 hours	21h
Virtual machine	Xeon 2.5GHz	32GB	Host: RAID50, 10k rpm	Linux	cancelled, predicted to take >160 hours	23-25h
Laptop	Core i7 2.4GHz	16GB	SSD + HDD	Windows	cancelled after 160 hours, infeasible	52h

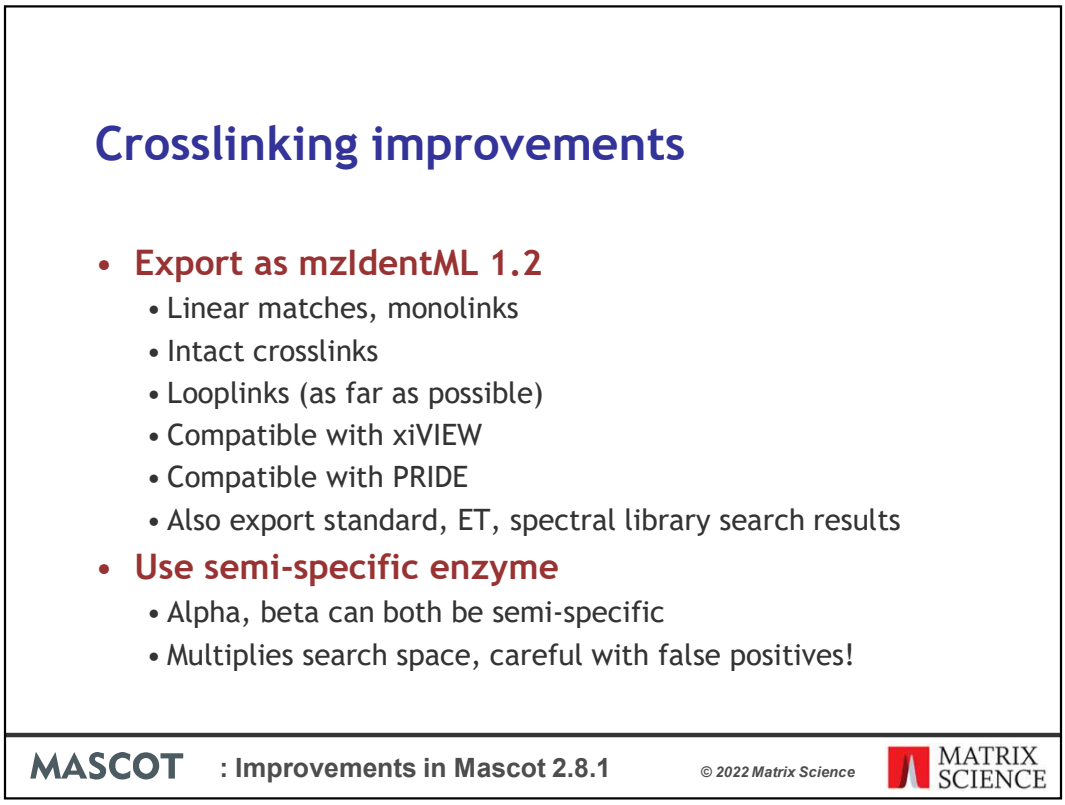
The second issue with a huge database like NCBIprot is compression speed. Mascot needs to create compressed index files before the database can be searched, and this had started to take absolutely forever on some systems. It turned out to be a disk bottleneck when parsing the taxonomy data. If you had a powerful RAID array and lots of RAM, the old version was OK. The slower the disk and the less RAM, the longer it took to compress NCBIprot. On medium to low-spec PCs and virtual machines, it could take several weeks.

We found a solution and the new version is a lot faster. The solution description is a bit technical, so if you're interested, please have a look at our April blog article.

We benchmarked the fixed version on a range of systems, and this table summarises the improvement. There is little difference on the server with a powerful RAID array. The last two columns have the timings for a database with 440 million sequences and another for 339 million sequences, so although it says 6 hours versus 5 hours, the new version is faster per sequence.

The new version is dramatically better on systems with less RAM and slower disk. The old version was infeasible on my laptop, the bottom row, and the new version finishes compression in a couple of days.

As the database continues to double in size, we recommend at least 32GB of RAM for now, and you'll probably want to upgrade to 64GB in a few years. A RAID array is very helpful too.



The patch release has two crosslinking improvements. You can now export crosslink searches as mzIdentML 1.2, in addition to CSV and XML. The file contains all peptide matches, monolinks and intact crosslinks. The format does not quite support looplinks, so these are currently exported as variable mods.

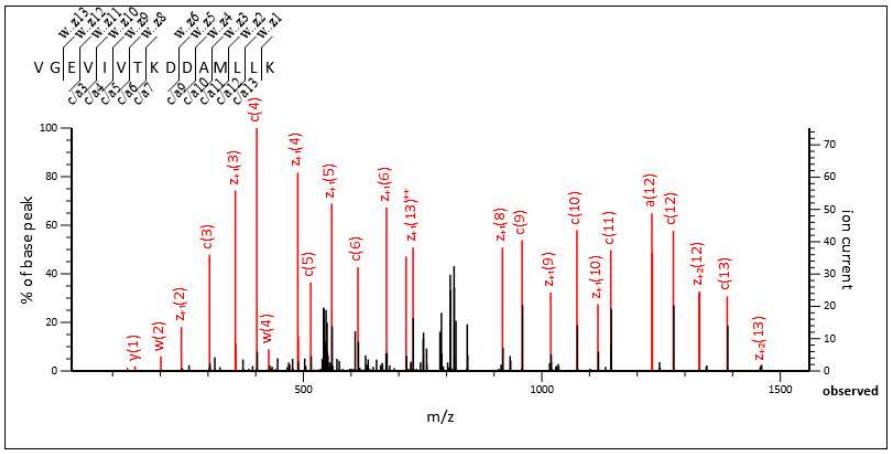
The mzIdentML export is fully compatible with xiVIEW. The PRIDE repository currently accepts mzIdentML 1.2 files but doesn't use the new metadata. Backend improvements for PRIDE repository are apparently planned to address this.

mzIdentML 1.2 isn't limited to crosslinked searches. You can now export any database search results in the new format.

We also lifted a restriction and you can now use a semi-specific enzyme in combination with intact crosslinking. Both alpha and beta peptides can be semi-specific. This will multiply the search space by a couple orders of magnitude, so you should review the matches carefully in case of false positives.

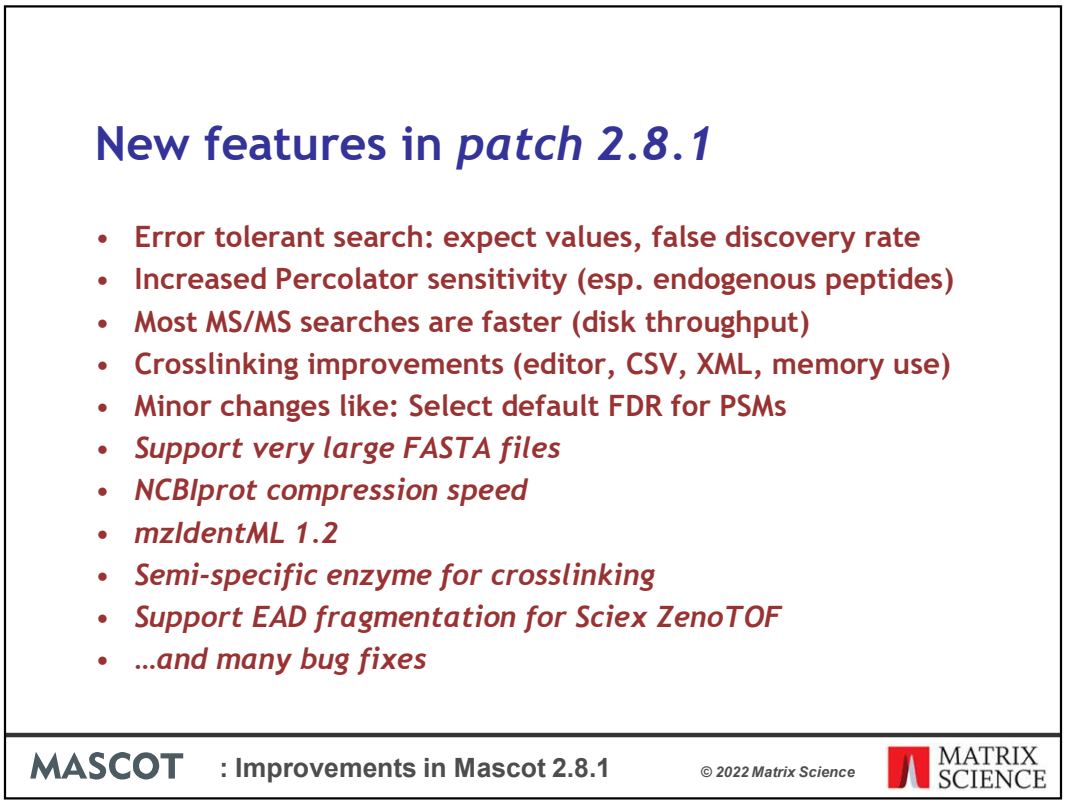
EAD fragmentation

- Sciex ZenoTOF, similar to EThcD



MASCOT : Improvements in Mascot 2.8.1 MATRIX SCIENCE

Another small improvement is an instrument definition for electron-assisted dissociation, EAD. Here's an example spectrum from a Sciex ZenoTOF. The peptide match has a range of ion series, including c, y, w, z+1 and z+2. It's similar to but not quite the same as EThcD.



There were also many smaller bug fixes. The full list is on our website.