SUPPORTING INFORMATION

Identification of Cleavable and Non-Cleavable Chemical Crosslinked Peptides with MetaMorpheus

Lei Lu[1], Robert J. Millikin[1], Stefan K. Solntsev[1], Zach Rolfs[1], Mark Scalf[1], Michael R. Shortreed[1], Lloyd M. Smith[1,2]*

Department of Chemistry[1], and Genome Center of Wisconsin[2], University of Wisconsin, Madison, Wisconsin 53706, United States

*Corresponding Author

Table of Contents

Supplementary Methods

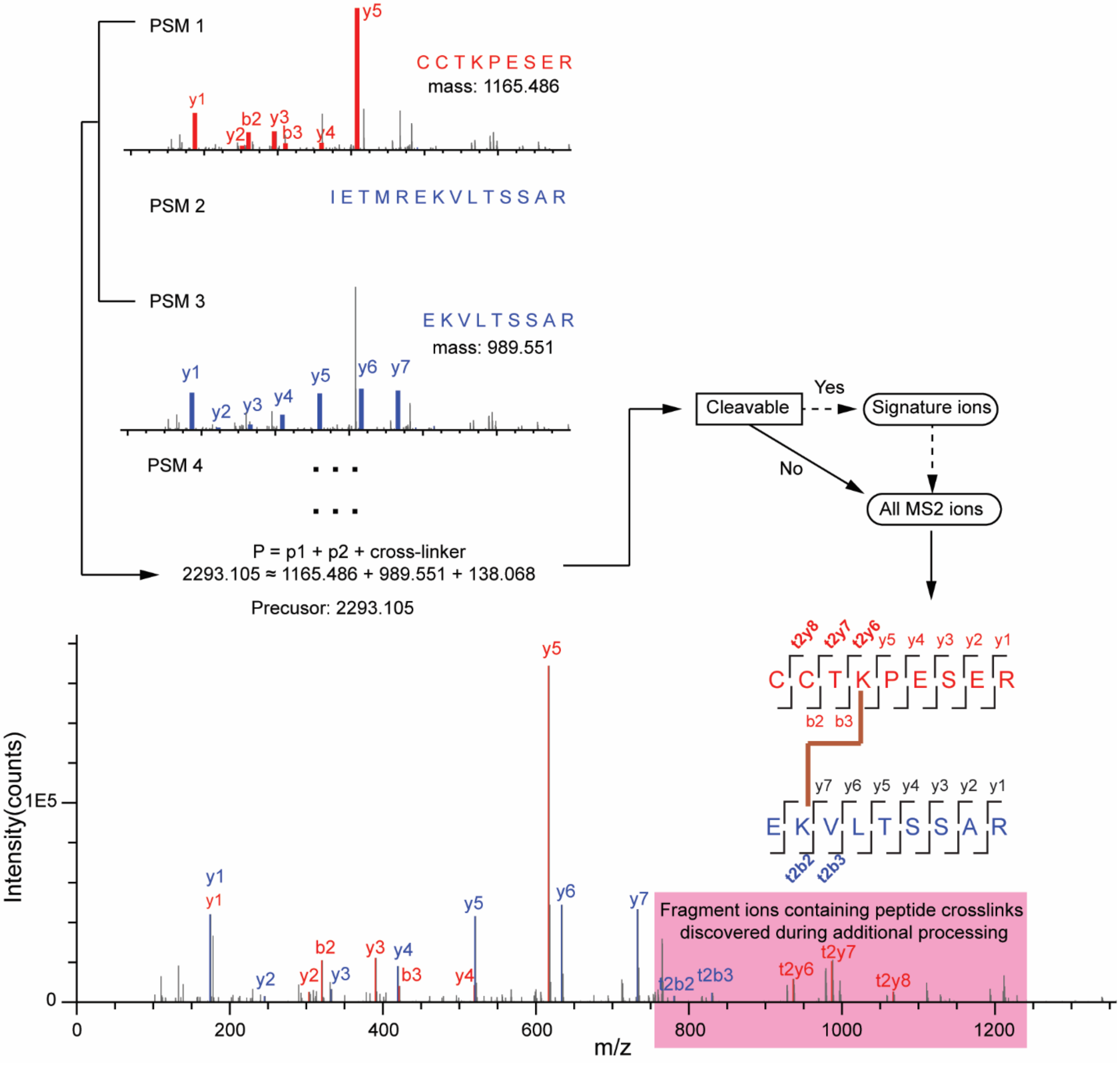
1. Ion-indexing: An “open-mass” search (i.e., when the precursor mass does not limit the space of theoretical peptides for fragment matching) with an ion-indexing strategy is used in MetaMorpheusXL. In this algorithm, the protein database is digested *in silico* and the digestion products (theoretical peptides) are written to a peptide index with each unique peptide being identified by an integer value (“ID”). The peptides are ordered by mass and each is fragmented *in silico*. For each theoretical fragment, its peptide’s ID is stored in a lookup table according to the fragment mass (rounded to the nearest mDa). Experimental fragments are matched to theoretical peptides by finding the experimental fragment’s mass in the lookup table. The peptide IDs in the fragment mass bin are filtered by the desired precursor mass tolerance (for completely open-mass searches, this tolerance is infinity and all peptides in the bin are counted as having matched to that experimental fragment ion).

2. Calibration: The mass accuracy of MS1 and MS2 spectra gathered during a proteomics experiment can vary significantly over the course of a single run and over the course of several runs. Systematic drift, random noise, and changes in temperature and other environmental conditions can contribute to this variation. Therefore, spectral mass calibration prior to the final analysis can improve peptide identification accuracy. MetaMorpheus uses a machine-learning algorithm to calibrate both MS1 and MS2 spectra. The process begins with a preliminary search of the uncalibrated file to identify a set of confident peptide spectral matches. Mass spectral peaks of confident PSMs are the calibration points, accompanied by several additional values, including: the difference between observed *m/z* and theoretical *m/z* (the “*m/z* error”), the absolute *m/z,* the retention time, the total ion current, and the ion injection time. All of these values serve as input to a random forest machine-learning algorithm that performs a regression analysis to model the *m/z* error as a function of the above explanatory variables. This function is used to shift the *m/z* of all peaks in all scans in the run. The calibrated spectra file is then used for a complete proteomics analysis.

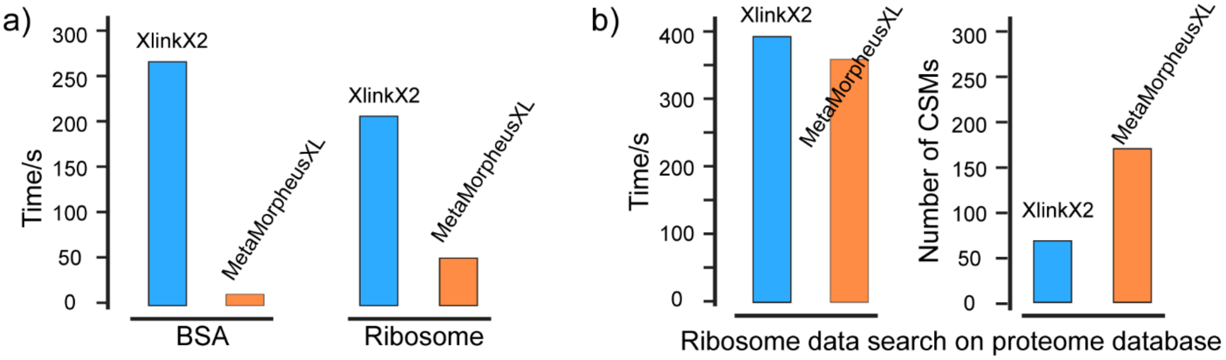
3. FDR estimation: The q-value for each CSM is determined by calculating the ratio of the count of CSMs assigned to target by the count of CSMs assigned to decoy with scores greater than or equal to the current CSM (q-value = (target count)/(decoy count)). In MetaMorpheusXL, a CSM is assigned as a target only when both peptides of the crosslink pair are present in the target database. When either member or both of a crosslink pair are present in the decoy database, the CSM is assigned as a decoy. This results in an imbalance in the total number of target and decoy pairs, which makes it more likely for a CSM to be assigned as a decoy than a target compared to a typical target-decoy search. Therefore, it is possible for the q-value to exceed 1.

4. Example explanation of MetaMorpheusXL's workflow: SI Figure S-1 An illustration of the algorithm used by MetaMorpheusXL to identify a peptide crosslink. This example uses a high-quality MS2 spectrum from a BSA sample crosslinked with DSS. The MS2 spectrum was obtained from a precursor species with a mass of 2293.105 Da. MetaMorpheusXL first finds all candidate peptides by matching primary fragment ions using an indexed-ion open search method (see the *Ion-Indexing* section of this supplement). All possible peptide matches are then paired with each other to generate candidates for crosslink pairs. A candidate pair is valid if the summed mass of the two peptides and the crosslink molecule matches the precursor mass. For example, PSM 1 and PSM 3 in the top panel of **SI Figure S-1** are considered a valid candidate pair because the summed mass of the two peptides and the crosslinker (1165.486 Da + 989.551 Da + 138.068 Da) are within the tolerance of precursor mass (2293.105 Da ± 10 ppm). Theoretical ions containing crosslinker-specific modifications are then generated for each crosslink candidate pair and matched to the spectrum; the highest-scoring (the CSM with the most matching fragments) is retained.

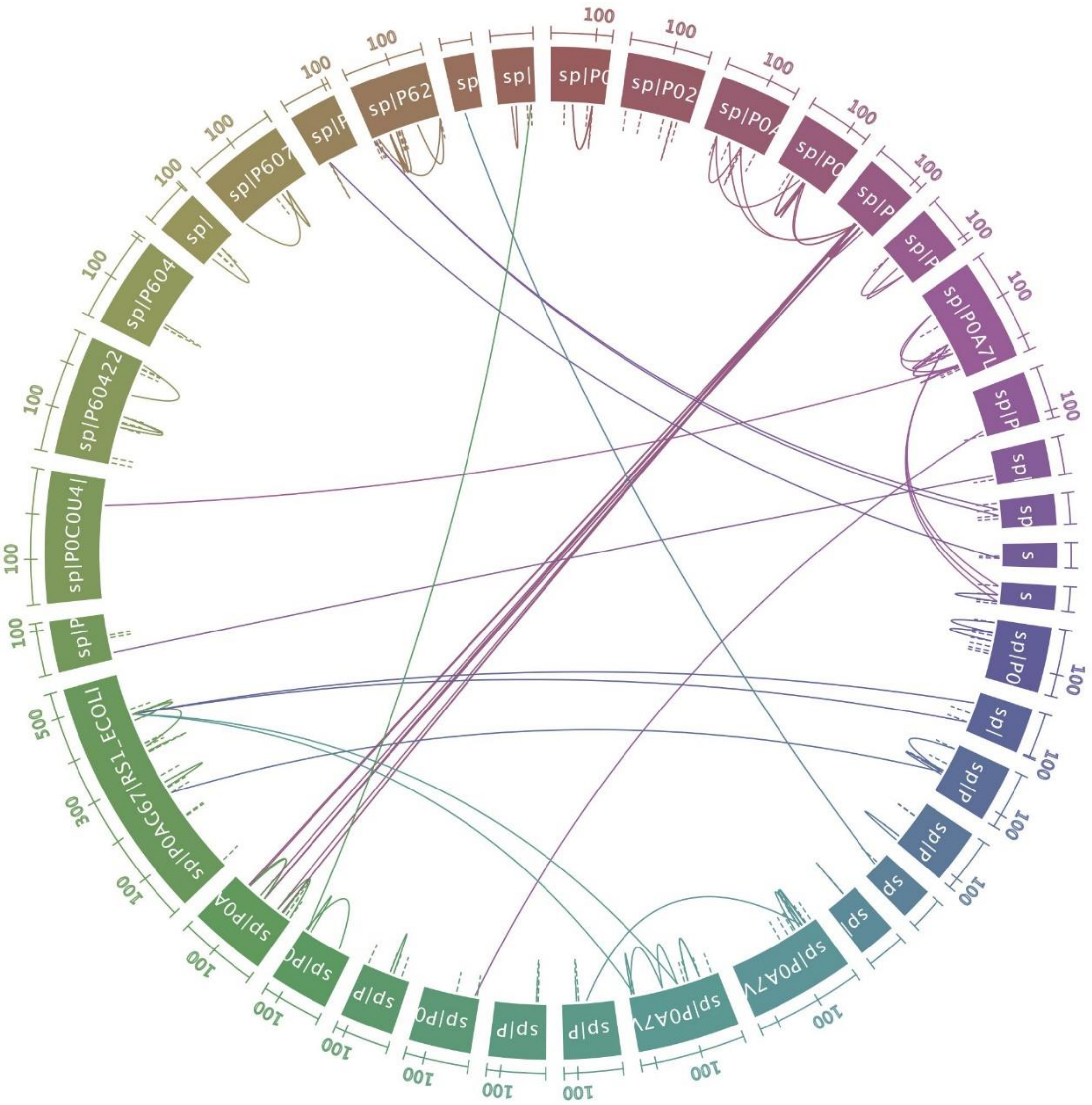
SI Figure S-1. Example of a crosslinked peptide identified by MetaMorpheusXL. The MS2 spectrum's precursor mass is 2293.105 Da. Preliminary peptide matches are generated with an open-mass search. The candidate PSM masses are paired, with a valid pair satisfying the equation $M_{precursor} = M_{alpha} + M_{beta} + M_{crosslinker}$. The highest-scoring pair that satisfies this constraint was PSM 1 (CCTKPESER) paired with PSM 3 (EKVLTSSAR). Fragment ions containing peptide crosslinks are discovered during an additional processing and the CSM score is increased by one for each additional fragment ion matches.



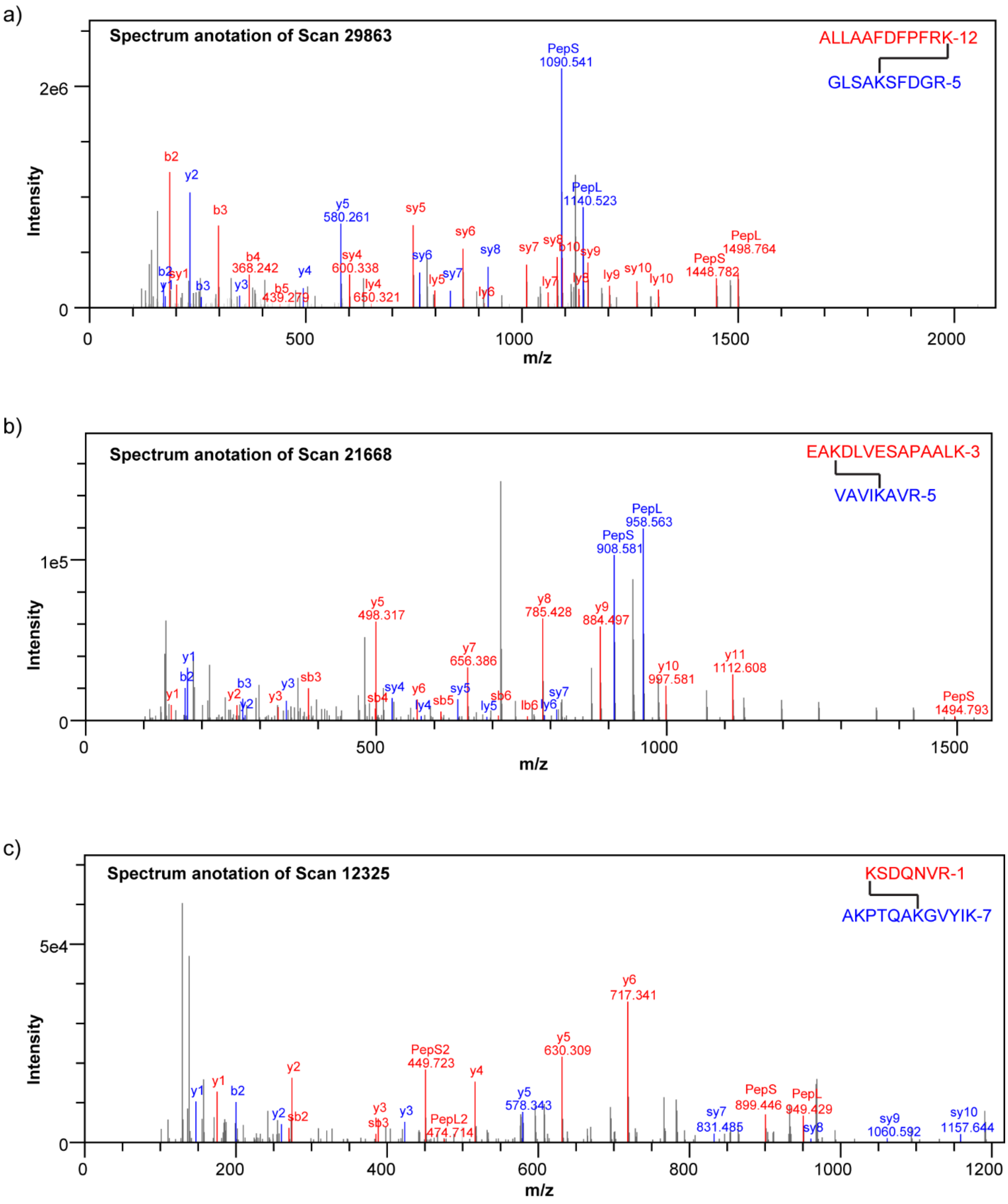
SI Figure S-2. Computation time comparison between MetaMorpheusXL and XlinkX 2.0. (a) Computation time comparison for searching BSA and ribosome data using small theoretical databases. (b) Computation times and numbers of CSMs identified when searching ribosome data against the entire *E. coli* proteome database, which contains 4443 proteins. MetaMorpheusXL took 6.4 min when restricted to the top 500 peptide matches per MS2 spectrum and identified 173 inter- and intra-CSMs combined, 35% less than the 262 inter- and intra-CSMs identified using the restricted database. Only 3 identified proteins were not ribosomal or ribosome-related. XlinkX 2.0 took 6.5 min and identified 66 CSMs, among which 21 proteins were not ribosomal or ribosome-related. Searches with MeroX and DXMSMS using the whole *E. coli* proteome database took too long to evaluate the results.



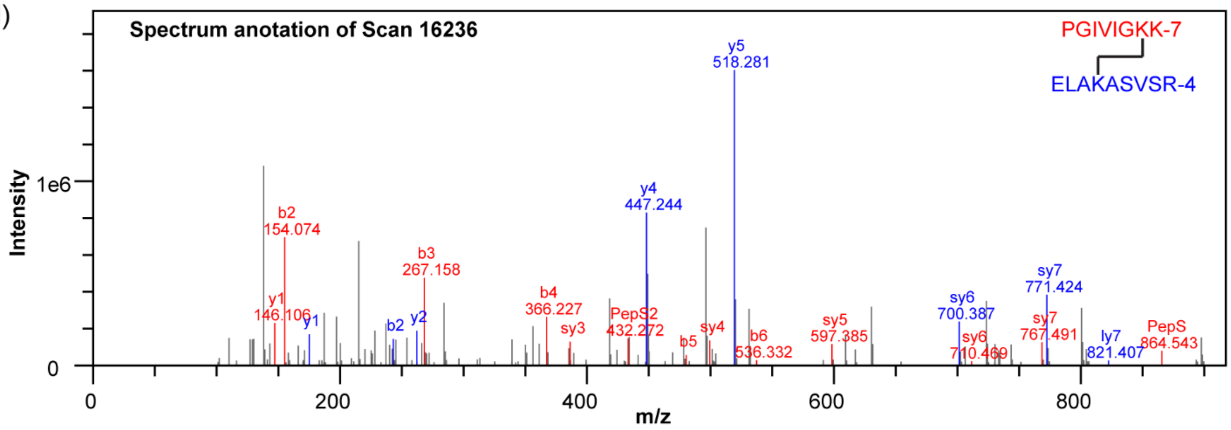
SI Figure S-3. Circle plot from ProXL displaying crosslinks of DSSO-crosslinked ribosome proteins. The bars represent proteins, lines represent crosslinks and dashed lines represent dead-ends.



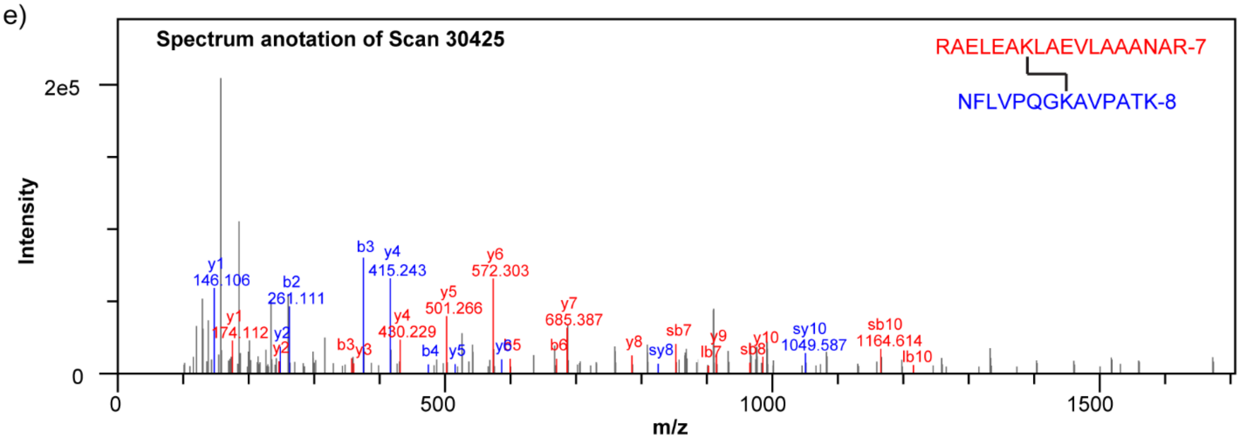
SI Figure S-4. Examples of annotated spectra from *E. coli* ribosome data with different numbers of identified signature ions (4 to 0) and from different score ranges (high to low). The “PepS” or “PepL” annotations indicate signature ions containing the short or long pieces of the fragmented MS-cleavable crosslinker molecule. “PepS2” indicates a doubly-charged signature ion with a short crosslinker piece. The “lb4” and “sb4” refer to b4 ions with long or short cleavage products, respectively. (a) This CSM contains 4 signature ions with a high score. (b) This CSM contains 3 signature ions. (c) This CSM contains 2 signature ions, both of which are from the alpha peptide. (d) This CSM contains 1 signature ion from its alpha peptide. (e) This CSM contains no signature ions. (f) This CSM contains no signature ions and is low-scoring.


a)
Spectrum anotation of Scan 29863
ALLAAFDFPFRK-12
GLSAKSFDGR-5
Intensity
2e6
0
m/z
0
500
1000
1500
2000
PepS
1090.541
PepL
1140.523
b2
y2
b3
y5
580.261
sy5
sy6
sy8
sy7
sy8
b10
sy9
PepL
1498.764
PepS
1448.782
b4
368.242
sy4
600.338
y4
b5
439.279
ly4
650.321
sy6
sy7
ly5
ly6
ly7
y8
ly9
sy10
ly10
b2
y1
sy1
b3
y3
b)
Spectrum anotation of Scan 21668
EAKDLVESAPAALK-3
VAVIKAVR-5
Intensity
1e5
0
m/z
0
500
1000
1500
PepL
958.563
PepS
908.581
y5
498.317
y8
785.428
y9
884.497
y7
656.386
y10
997.581
y11
1112.608
PepS
1494.793
y1
b2
y1
y2
y2
b3
y3
y3
sb3
sb4
sy4
y6
ly4
sb5
sy5
ly5
sb6
lb6
ly6
sy7
c)
Spectrum anotation of Scan 12325
KSDQNVR-1
AKPTQAKGVYIK-7
Intensity
5e4
0
m/z
0
200
400
600
800
1000
1200
y6
717.341
y5
630.309
PepS2
449.723
y4
y2
y1
y1
b2
y2
sb2
y3
sb3
y3
PepL2
474.714
y5
578.343
PepS
899.446
PepL
949.429
sy7
831.485
sy8
sy9
1060.592
sy10
1157.644

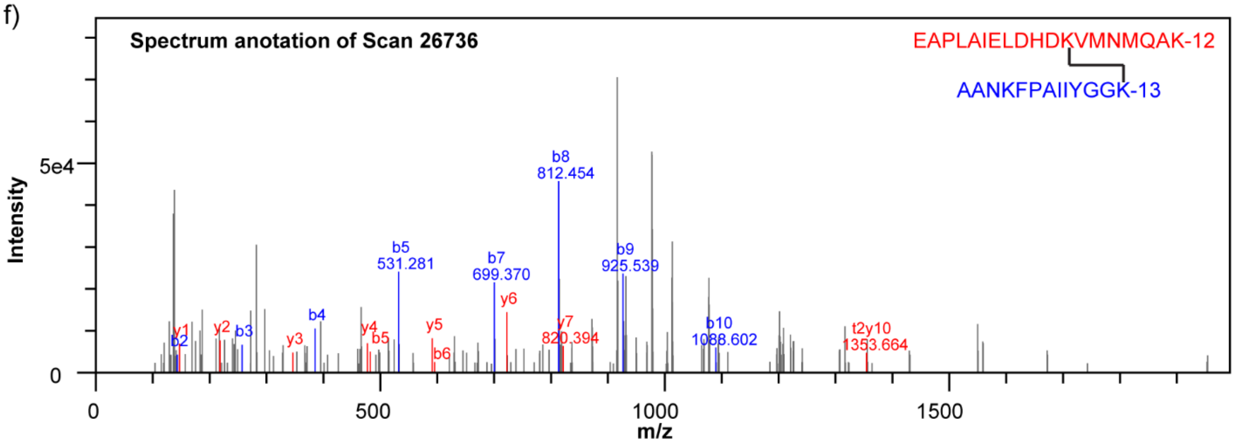
d)


Spectrum anotation of Scan 16236
PGIVIGKK-7
ELAKASVSR-4
Intensity
1e6
0
m/z
0
200
400
600
800
y5
518.281
y4
447.244
b2
154.074
b3
267.158
y1
146.106
y1
b2
y2
b4
366.227
sy3
PepS2
432.272
b5
sy4
b6
536.332
sy5
597.385
sy6
700.387
sy6
710.469
sy7
771.424
sy7
767.491
ly7
821.407
PepS
864.543

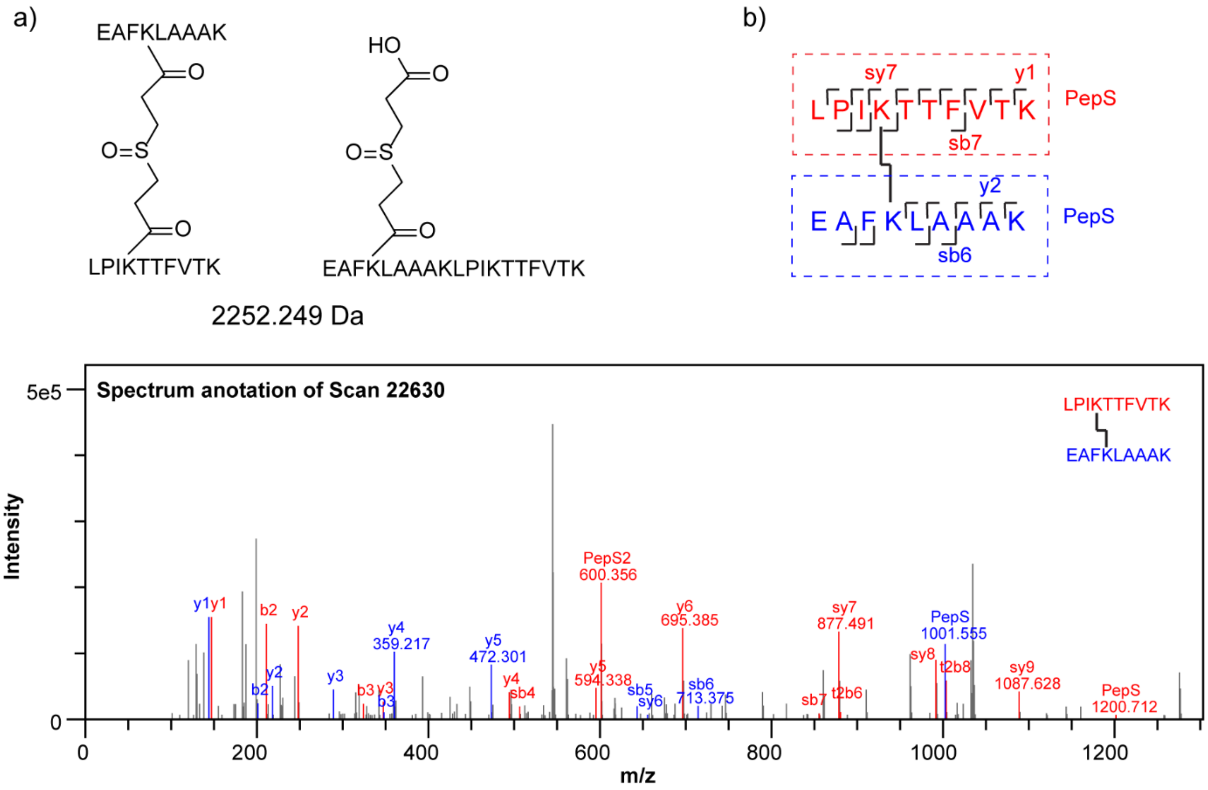
e)


Spectrum anotation of Scan 30425
RAELEAKLAEVLAAANAR-7
NFLVPQGKAVPATK-8
Intensity
2e5
0
m/z
0
500
1000
1500
y1
146.106
b2
261.111
b3
y4
415.243
y6
572.303
y5
501.266
y4
430.229
y1
174.112
y2
y2
b3
y3
b4
y5
y6
b5
b6
y7
685.387
y8
sy8
sb7
lb7
y9
sb8
y10
sy10
1049.587
sb10
1164.614
lb10

f)


Spectrum anotation of Scan 26736
EAPLAIELDHDKVMNMQAK-12
AANKFPAIIYGGK-13
Intensity
5e4
0
m/z
0
500
1000
1500
b8
812.454
b5
531.281
b7
699.370
b9
925.539
y6
b4
y1
b2
y2
b3
y3
y4
b5
y5
b6
y7
820.394
b10
1088.602
t2y10
1353.664

SI Figure S-5. Identification of intra-protein crosslinks composed of consecutive sequences. (a) Intra-crosslinks composed of consecutive sequences (left) have the same precursor mass as the dead-end missed-cleavage product modified with hydrolyzed crosslinker (right). (b) MetaMorpheusXL assigns a crosslink composed of consecutive sequences as an intra-crosslink only if the matched fragment ions could differentiate it from the dead-end crosslink. From the ribosome data, an intra-crosslink composed of consecutive sequences 'EAFKLAAAK' and 'LPIKTTFVTK' of protein P0ADY7 are shown as an example here. The spectral matches containing indicative fragment ions (*e.g.*, the y2 ion of 'EAFKLAAAK') support that the pair is an intra-crosslink instead of dead-end missed-cleavage product.



SI Table-1. Parameters used in this work for searches of crosslinked data with MetaMorpheusXL, XLinkX 2.0 and Kojak 1.5.

MetaMorpheusXL parameters

Crosslinker type: The crosslinker molecule used in the sample; can be user defined.

Search top matches: if selected, MetaMorpheusXL will only consider N top-scoring peptides for peptide pairing.

Search top Num: used together with 'Search top matches'; this defines the N top peptide candidates.

Trim MS/MS peaks: only match the most intense peaks in an MS2 spectrum. Used together with 'Top N peaks' (i.e., N most intense peaks) and 'Minimum ratio' (peaks must be this intense compared to the base peak).

Minimum Score allowed: the lowest peptide score after the 'first pass' that will be considered.

parameters	Cleavable	Non-cleavable
Precursor Mass tolerance	10 ppm	10 ppm
Crosslinker type	DSSO	DSS
Search top matches	-	✓
Search top Num	-	300
Use Provided Precursor	✓	✓
Deconvolute Precursor	✓	✓
Trim MS1 Peaks	-	-
Trim MS/MS Peaks	✓	✓
Top N Peaks	200	500
Minimum ratio	0.01	0.005
Generate decoy proteins	✓	✓
Max missed cleavages	2	2
protease	trypsin	trypsin
Initiator methionine	Variable	Variable
Max modification isoforms	4096	4096
Min peptide length	5	5
Product mass tolerance	20 ppm	20 ppm
Ions to search	B ions, Y ions	B ions, Y ions
Minimum Score allowed	5	2
Fixed modification	Carbamidomethyl of C	Carbamidomethyl of C
Variable modification	Oxidation of M	Oxidation of M
Localize all modification	✓	✓
Output for Percolator	-	✓

Output for Crosslink	✓	✓

Parameters used in XlinkX2.0

XlinkX 2.0 was used as a node in Thermo Scientific Proteome Discoverer 2.2. Parameters used in XlinkX 2.0 are listed below.

XlinkX 2.0 Detection
Acquisition strategy: MS2
Crosslink Modification: DSSO / + 158.004 Da (K)
Minimum S/N: 1.5
Enable protein N-terminus linkage: false
Xlinkx Filter
Select: Crosslinks
Xlinkx Search
Retain FASTA file indexes: True
Enzyme Name: Trypsin(full)
Maximum Missed Cleavages: 2
Maximum Peptides Considered: 10
Minimum Peptide Length: 5
Maximum Number Modifications: 3
Minimum Peptide Mass: 300
Maximum Peptide Mass: 7000
Precursor Mass Tolerance :10 ppm
FTMS Fragment Mass Tolerance: 20 ppm
Static Modification: Carbamidomethyl / + 57.021 Da (C)
Dynamic Modification: Oxidation /+ 15.995 Da (M)
FDR threshold: 0.01
FDR strategy: Percolator

3.3 Parameter used in Kojak1.5

```
percolator_version = 3.0
enrichment = 0
instrument = 0
MS1_centroid = 1
MS2_centroid = 1
MS1_resolution  = 100000
MS2_resolution  = 7500
cross_link = nK         nK 138.0680742 DSS
mono_link = nK 156.0786
fixed_modification = C 57.02146
fixed_modification_protC = 0
fixed_modification_protN = 0
modification = M 15.9949
modification_protC = 0
modification_protN = 0
diff_mods_on_xl = 0
```

```
max_mods_per_peptide = 2
mono_links_on_xl = 0
enzyme = [KR]|{P}
fragment_bin_offset = 0.0
fragment_bin_size = 0.03
ion_series_A = 0
ion_series_B = 1
ion_series_C = 0
ion_series_X = 0
ion_series_Y = 1
ion_series_Z = 0
decoy_filter = DECOY
isotope_error = 1
max_miscleavages = 2
max_peptide_mass = 8000.0
min_peptide_mass = 500.0
max_spectrum_peaks = 0
ppm_tolerance_pre = 10.0
prefer_precursor_pred = 2
spectrum_processing = 0
top_count = 300
truncate_prot_names = 0
turbo_button = 1
```

MetaMorpheusXL User Manual

1. Download the current version of MetaMorpheus from https://github.com/smith-chem-wisc/MetaMorpheus/releases. MetaMorpheusInstaller.msi is suggested for Windows users.
2. Double-click the .msi file to install MetaMorpheus. Open MetaMorpheus after installation.
3. Click the ‘New XL Task’. This will open a window to set the parameters for a new crosslink search. Parameters are described in this document, below. After choosing your parameters, click ‘Add the XLSearch Task’.

   Crosslink Search panel:

   - ‘Crosslink Precursor mass tolerance’: Sets precursor mass tolerance, in Daltons (Da) or parts per million (ppm).
   - ‘Crosslinker Type’: choose the crosslinker used in your sample. If ‘UserDefined’ is chosen, additional crosslinker information must be specified.
   - ‘Search Top matches’: this option can help speed up searches. This option defines the number of peptides from the open search to pair.

   Search Parameters panel:

   - ‘Ions to search’: Ion types should be specified to match the fragmentation method (e.g., b and y ions for HCD data).
   - When searching a whole proteome database, selecting ‘Search Top matches’ is recommended, along with setting the number of database partitions to ~4-8.

4. Drag your database and spectra files into MetaMorpheus and click ‘Run All tasks’. The results will be in the same folder as the data files.
5. If you have any problems, support is available by reading the wiki (Help -> Open Wiki page), opening an issue on GitHub (Help -> Submit an issue on GitHub), or by emailing the MetaMorpheus development team at mm_support@chem.wisc.edu .