Supplemental Material

OpenPepXL: An open-source tool for sensitive identification of cross-linked peptides in XL-MS

Eugen Netz[1,2,3*], Tjeerd Dijkstra[1,2,3,4], Timo Sachsenberg[2,3], Lukas Zimmermann[2,5], Mathias Walzer[6], Thomas Monecke[7,8], Ralf Ficner[8], Olexandr Dybkov[9], Henning Urlaub[10,11], Oliver Kohlbacher[1,2,3,5,12]

[1]Biomolecular Interactions, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.
[2]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany.
[3]Applied Bioinformatics, Dept. of Computer Science, University of Tübingen, 72076 Tübingen, Germany.
[4]Center for Women's Health, University Clinic Tübingen, 72076, Tübingen, Germany.
[5]Institute for Translational Bioinformatics, University Hospital Tübingen, 72076 Tübingen, Germany.
[6]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
[7]X-Ray Crystallography Facility, Institute of Pharmaceutical Biotechnology, University of Ulm, 89081 Ulm, Germany.
[8]Department of Molecular Structural Biology, Institute for Microbiology and Genetics, GZMB, Georg-August-University Goettingen, 37077 Goettingen, Germany.
[9]Department for Cellular Biochemistry, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany.
[10]Bioanalytical Mass Spectrometry, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany.
[11]Bioanalytics, Institute for Clinical Chemistry, University Medical Center, 37075 Göttingen, Germany.
[12]Quantitative Biology Center, University of Tübingen, 72076 Tübingen, Germany.

Supplemental methods

Mass spectrometry of CRM Complex

Human full-length wild type CRM1 and SNP1 as well as a 1-180 fragment of Ran carrying a Q69L mutation, which blocks GTPase activity, were recombinantly expressed in E. coli, purified and a trimeric complex thereof was assembled as described before[1]. Prior to cross-linking, the trimeric complex was dialyzed against a buffer containing 20 mM HEPES-KOH, pH 7.9, 50 mM NaCl, 2 mM $Mg(CH_3COO)_2$, 1 mM DTT. One hundred pmol of the trimeric complex was cross-linked with 150 µM bis(sulfosuccinimidyl)suberate (BS3, Thermo Fisher Scientific) at 25 °C for 30 min in a volume of 50 µl and subsequently quenched for 5 min with 0.1 M Tris-HCl, pH 8. The sample was resolved on a NuPAGE 4-12% Bis-Tris protein gel (Thermo Fisher Scientific), followed by Coomassie staining. A slow migrating band corresponding to the BS3-cross-linked products was excised and in-gel digested with trypsin. Extracted peptides were dissolved in a sample solvent (5% acetonitrile and 0.1% formic acid). Approximately 5 pmol peptides were injected into an EASY-nLC 1000 HPLC system coupled to Q Exactive mass spectrometer (Thermo Fisher Scientific) in duplicates under each of the three tested normalized collision energy (NCE) conditions using a 50-min method. A 20 cm long C18 analytical column with inner diameter of 75 µm self-packed with 5 µm beads (pore 120 Å, Dr. Maisch) was used for on-line HPLC. MS1 and MS2 resolution were set to 70000 and 17500, respectively. Fifteen most abundant precursors with charge of 3-7 (350-1600 m/z, isolation window 2 m/z) were selected for MS2 fragmentation at NCE 20, 24 or 28%. MS2 AGC target and injection time were limited to 200000 and 60 ms, respectively. Dynamic exclusion of 15 s was applied.

Implementation of spectrum pre-processing

The analysis starts with reading in the spectra and the protein database. The spectra are deisotoped and filtered by keeping only the 20 highest intensity peaks in a jumping window of 100 along the m/z axis. This will keep at most 400 peaks below 2000 m/z. The charge information obtained by deisotoping is stored for later use. Spectra that contain less than two times the minimal peptide size of peaks (10 peaks with the standard setting of minimum 5 residues per peptide) after these steps are not analyzed further. Additional preprocessing of the spectra differs between the workflows for labeled and label-free cross-linkers. For labeled cross-linkers an additional consensusXML file produced by the TOPP tool FeatureFinderMultiplex is read in that contains a pairing of MS1 features. These are used to link together MS2 spectra with the light and heavy versions of isotopically labeled cross-links. In the case of multiple MS2 spectra per MS1 feature, all possible combinations between one light and one heavy MS2 spectrum from these two features are considered as separate pairs. In the end the top ranking peptide pair candidates to all these pairs are summarized to one ranked list per light spectrum. The MS2 spectrum pairs are combined by first matching peaks between them and storing these as a linear ion spectrum. Then peaks shifted by the mass of the isotopic label for all considered charges are matched and the corresponding peaks from the light spectrum are stored as a cross-linked ion spectrum. Here, if the charge determined from deisotoping the spectrum does not match the charge considered for the mass shift, the matched peak is not accepted as a cross-linked ion. The intensity ratio of the peaks is also considered and cannot be lower than 0.3 for the lower intensity divided by the higher intensity. For the peaks whose charge could not be determined by deisotoping, the charge of this peak match is set as the known charge for matching to theoretical spectra later. This way charge determination from multiple sources is combined and used to filter out false positive peak matches, since we know the expected charge for every theoretical peak. For the label-free algorithm only the

charge information from deisotoping can be considered.

Implementation of candidate generation and scoring

The protein database is digested into modified peptides according to the enzyme and modification settings. The peptides are stored in a list sorted by their precomputed masses. Then the enumeration of candidates and scoring using the main score are done for each MS2 spectrum pair (labeled linkers) or MS2 spectrum (label-free) in a loop. Using the precomputed peptide masses, all peptide pairs whose mass including the cross-linker mass fits the MS2 spectrum's precursor mass tolerance window are enumerated and candidates with all possible cross-linked residue pairs are generated. To save memory, the peptide pair data structure stores indices from the list of linear peptides. This way the sequence and modification state of each peptide is only stored once even if it is reused in multiple peptide pair candidates. Candidate peptides with mono-links and loop-links are also considered. At this point a parallelized loop starts and scores each candidate for the current MS2 spectrum independently from the others. The scoring consists of generating theoretical spectra, matching theoretical and experimental spectra and computing the score. Four separate theoretical spectra are computed as needed, the spectrum containing linear ions from alpha peptide fragmentation, the spectrum containing cross-linked ions from alpha peptide fragmentation and the same two spectra for the beta peptide. For labeled cross-linkers, the linear theoretical spectra are matched to the linear ion spectrum and the cross-linked theoretical spectra are matched to the cross-linked ion spectrum. For label-free cross-linkers all four theoretical spectra are matched to the whole ion spectrum. The match-odds score is computed based on the number of matched peaks for each of these matches and the average of all four is the total match-odds score. The precursor mass error in ppm between the experimental precursor mass and the theoretical precursor mass of the candidate is computed as well. These values are combined into the final score (Formula 2). This formula was determined by an agreement between a linear

regression and a linear discriminant analysis done to find the best linear combination of many scores we tested, to separate target from decoy hits on several XL-MS datasets. The hits are sorted by this score and the top X (for a user defined X) hits are kept. For these top X hits more information, including fragment annotations for visualization and additional match quality metrics are computed afterwards.

For FDR calculation we implemented the formula from xProphet[2] into the OpenMS tool XFDR. The algorithm separates the hits into intra- and inter-protein cross-links, as well as an additional class for mono-links (or dead-end links) and loop-links. Hits to the target and decoy versions of the same protein are considered intra-protein cross-links. The FDR for the two cross-link classes is calculated separately using the xProphet[2] formula with the counts of target-target (*TT*), decoy-decoy (*DD*), target-decoy (*TD*) and decoy-target(*DT*) hits:

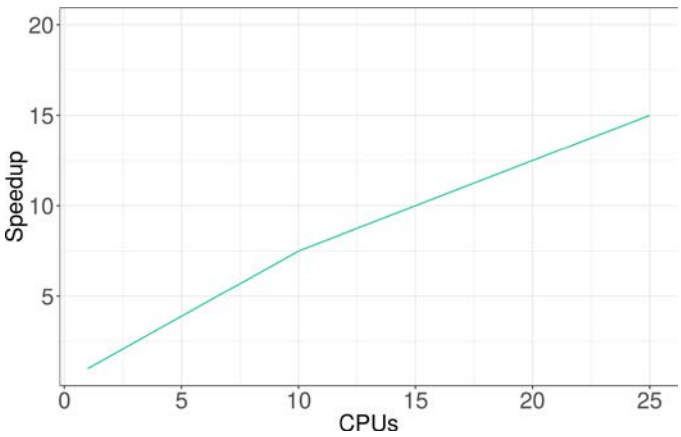
$$\frac{(TD + DT + DD) - 2 * DD}{TT} \qquad (S1)$$

Implementation of the peak matching algorithm:

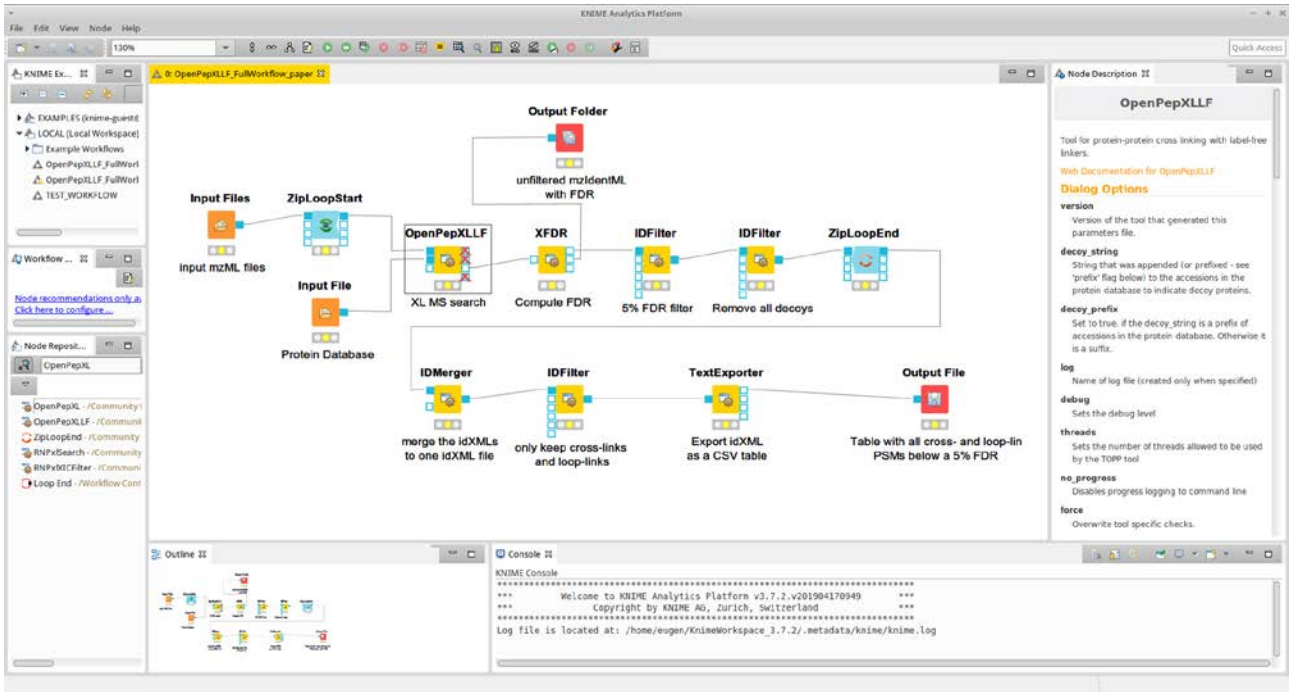
The peak matching algorithm to match two spectra is a linear sliding window algorithm that moves along one spectrum and searches for the best match in the corresponding tolerance window of the other spectrum. It is greedy in the sense that the peak from the second spectrum is the optimal match to the considered peak from the first spectrum, but this peak from the first spectrum might not be the best possible match for the matched peak from the second spectrum. It will always find the same number of matches as an optimal algorithm, but in ambiguous cases with multiple peaks within the tolerance window the matched pairs might be different if the input spectra are switched. In any case, the algorithm does not match peaks with disagreeing charges, if both charges are known and

an intensity ratio cut-off between matched peaks can also be considered. This algorithm is used both for matching light and heavy spectra from labeled cross-linker experiments and for matching experimental to theoretical spectra for the scoring of hits.

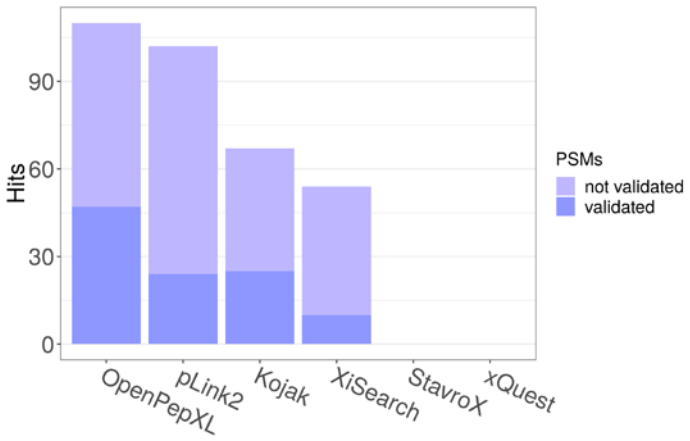
Supplemental Figures



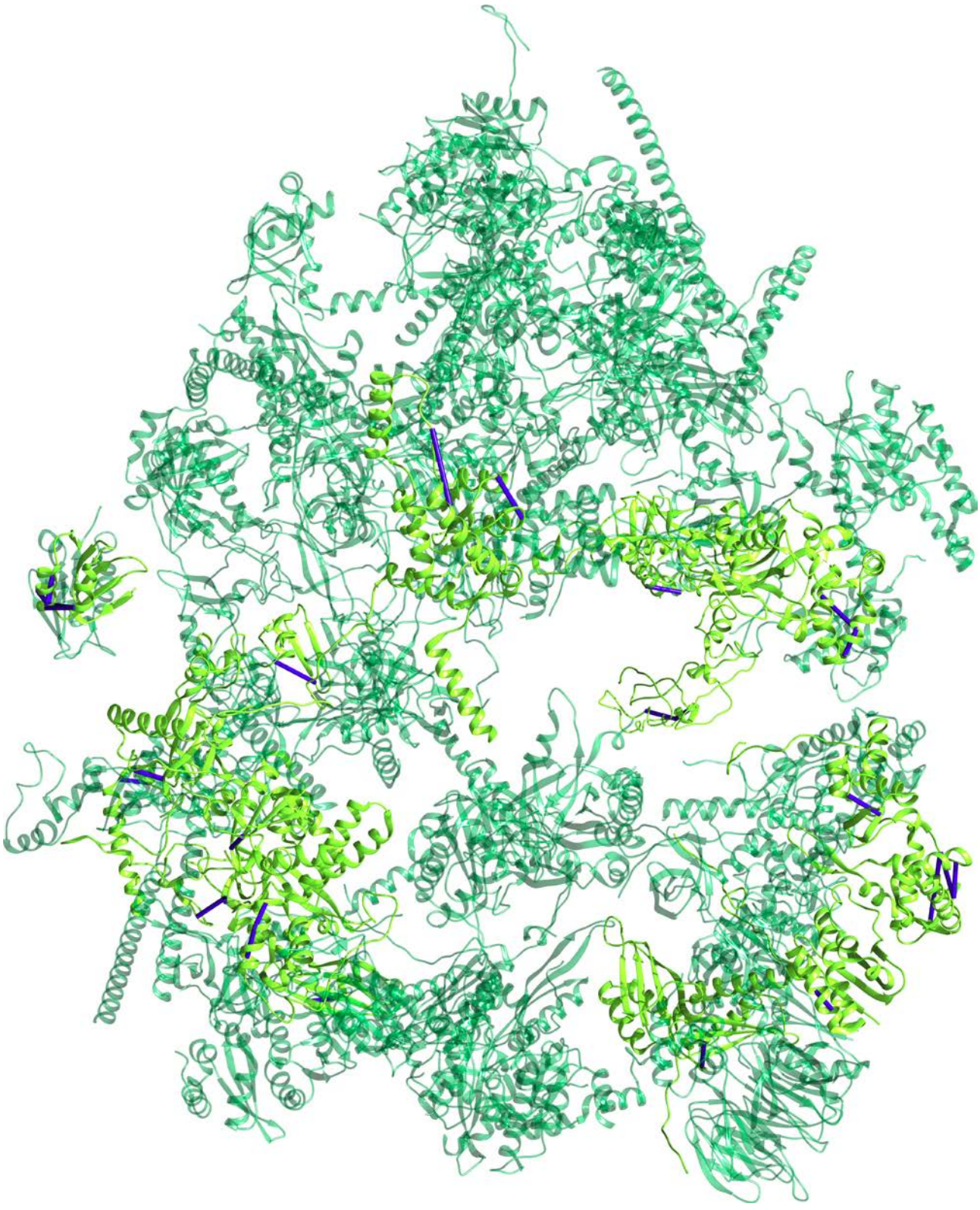
Supplemental Figure S1: Speedup of the OpenPepXL algorithm when using up to 25 CPU cores relative to the speed of using one core. When using 25 CPU cores to run one analysis in parallel, OpenPepXL runs 15 times faster than on one core and uses almost the same amount of memory. The runtime was measured with 1, 10 and 25 CPU cores on the same dataset and computer.



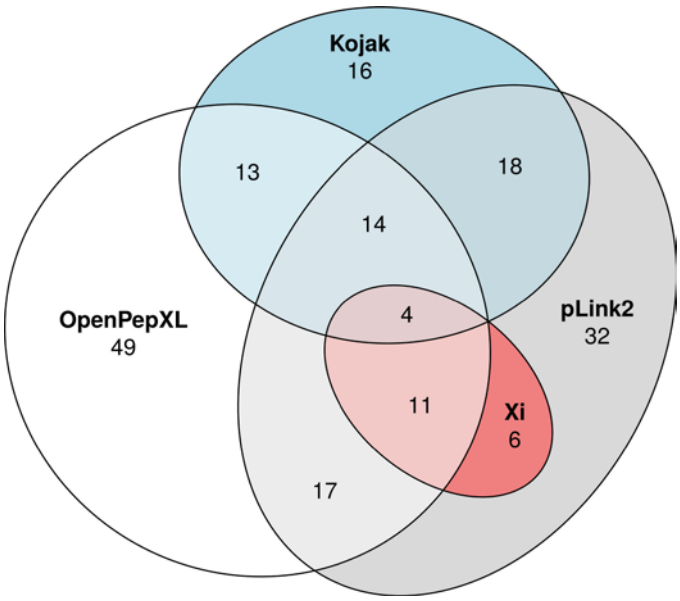
Supplemental Figure S2: The KNIME user interface showing a workflow for a label-free XL-MS analysis with OpenPepXL with FDR estimation and filtering of results. For information on how to install and get started with OpenPepXL in KNIME, visit https://www.openms.de/openpepxl/



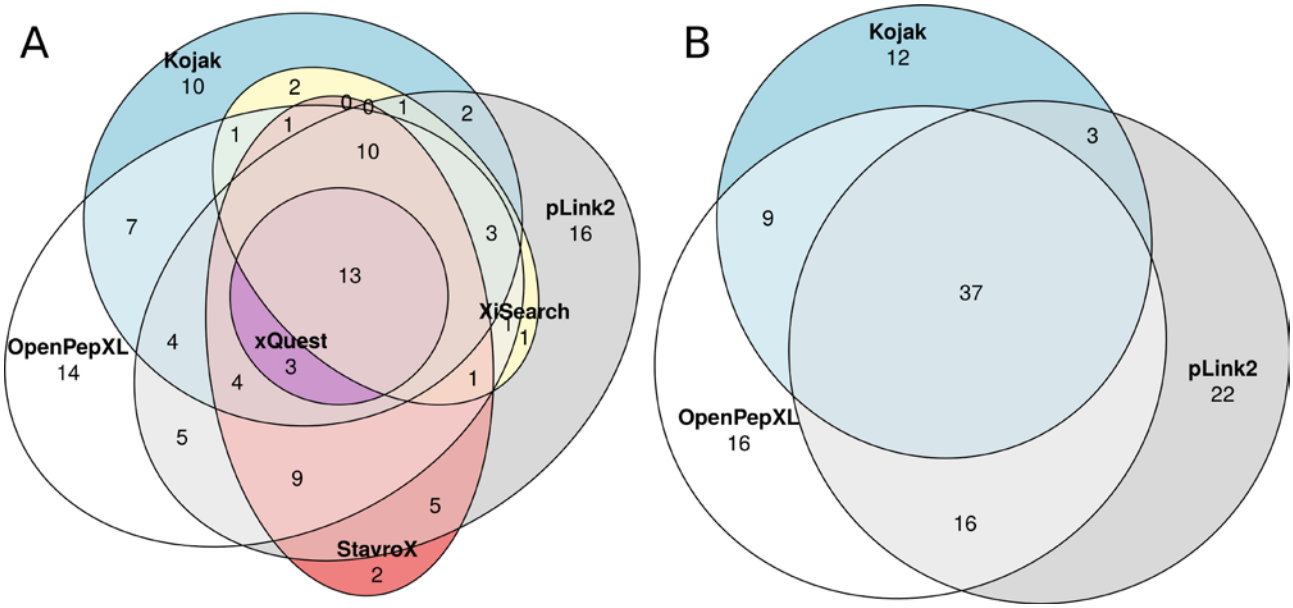
Supplemental Figure S3: Structurally validated unique residue pairs (URPs) from the ribosomal fraction data set. The validated URPs were covered by the structures we used for validation and none of them has a solvent accessible surface distance of more than 35 Å between C_β atoms. The rest of the URPs were not covered by currently published protein structures. Full list of used structures (UniProt sequence ID : PDB ID): P05387:2W1O, P06748:2LLH, P06748:2P1B, P11142:4H5R, P12268:1B3O, P14868:4J15, P19338:2KRR, P19338:2FC9, P21333:3CNK, P23396:5AJ0, P27635:5AJ0, P30050:5AJ0, P39019:5AJ0, P54136:4R3Z, P56192:5GL7, P61353:5AJ0, P62081:5AJ0, P62249:5AJ0, P62277:5AJ0, P62280:5AJ0, P62424:5AJ0, P62701:5A2Q, P62847:5AJ0, P62851:5AJ0, P62888:5AJ0, P62906:5AJ0, P62906:5AJ0, P63173:5AJ0, P63244:5AJ0, P67809:5YTT, P83731:5AJ0, Q00610:2XZG, Q12904:1FL0, Q14152:3J8B, Q53YD7:5JPO, Q5U0F4:5K0Y, Q6PIN5:3J2I, Q7L2H7:3J8B



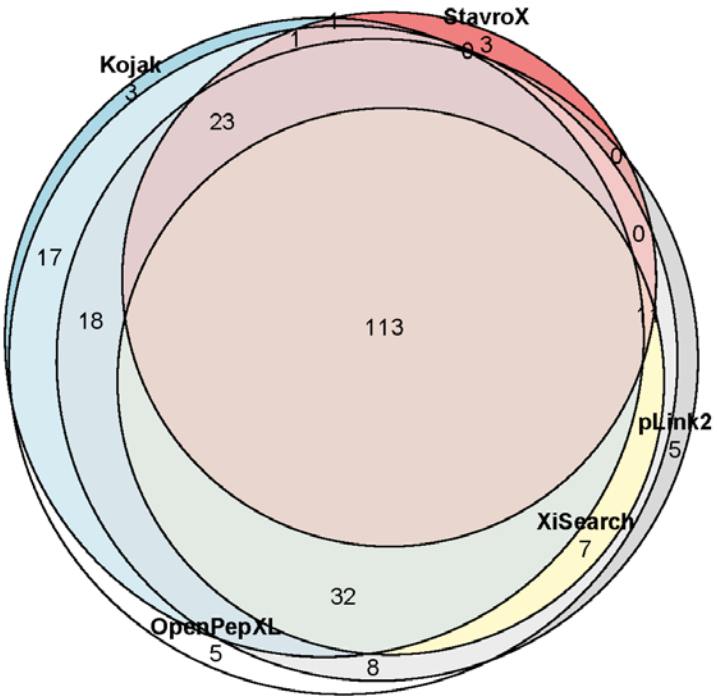
Supplemental Figure S4: Cross-links identified by OpenPepXL mapped onto the human ribosome structure (PDB ID: 5AJ0). The RNA was removed from the structure for visual clarity. Proteins without cross-links are shown in dark green, proteins with cross-links are shown in a lighter green. All cross-links are visualized as blue bars.



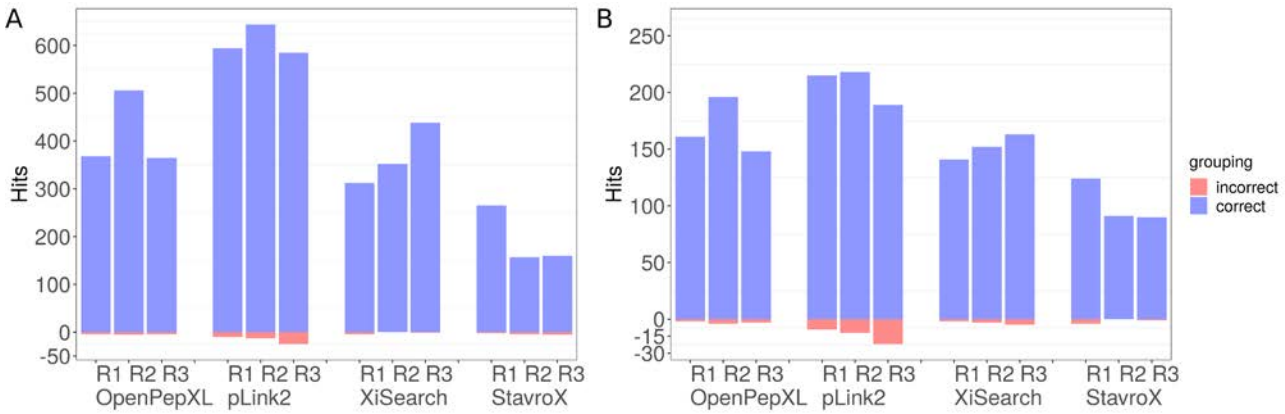
Supplemental Figure S5: Venn-Diagram of identified URPs for the ribosomal fraction dataset



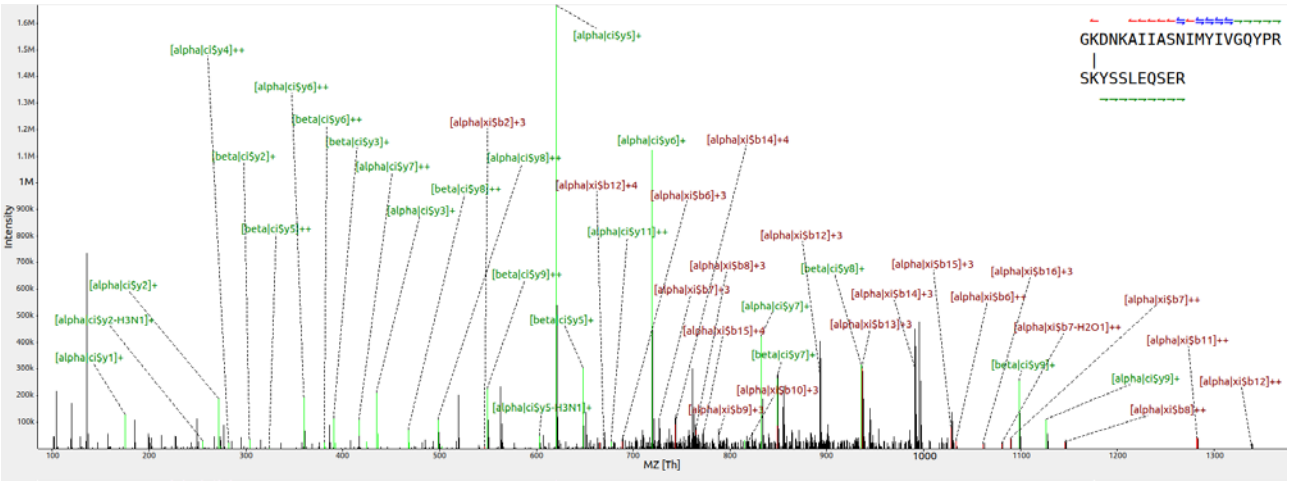
Supplemental Figure S6: Venn-Diagram of identified URPs for the CRM complex dataset. A) Includes all tools. B) Only includes the three tools with the highest number of identifications for visual clarity.



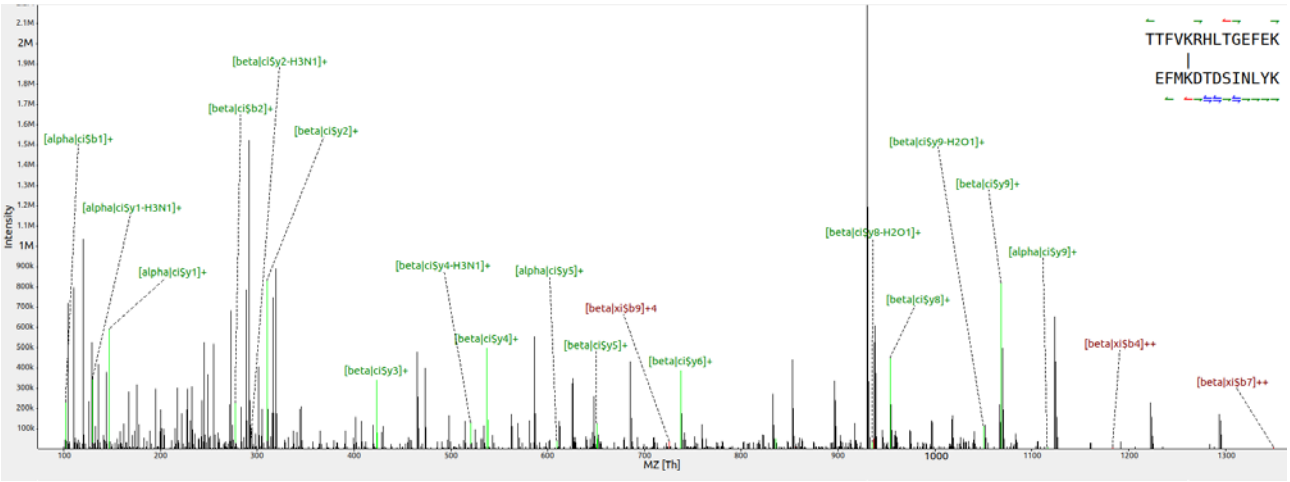
Supplemental Figure S7: Venn-Diagram of identified URPs for the synthetic peptides dataset.



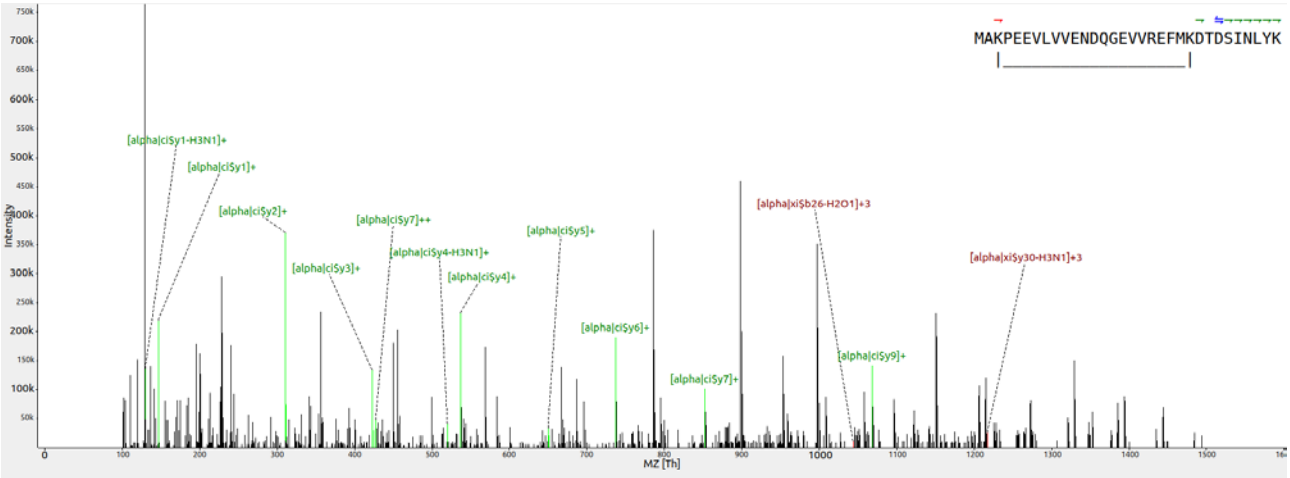
Supplemental Figure S8: The results from the search of the synthetic peptides dataset at a 1% FDR cut-off. All three replicates R1, R2 and R3 are shown. The blue bars show the number of valid CSMs/cross-links and the red bars on the negative Y-axis show the number of false positive identifications. All data except for OpenPepXL was taken from Beveridge *et al*[28]. xQuest was omitted, because it was not considered in that publication. Kojak was omitted, because the 1% FDR results were not available in that publication. (A) Number of reported CSMs. The exact numbers are in Supplementary Table S4. (B) Number of identified URPs. The exact numbers are in Supplementary Table S5.



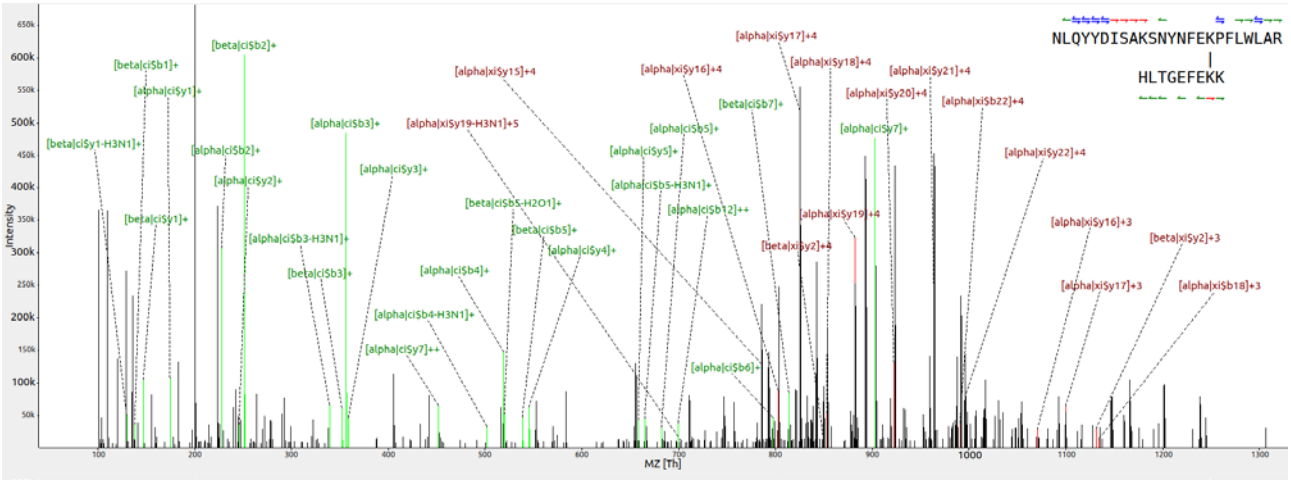
Supplemental Figure S9: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This URP was structurally validated.



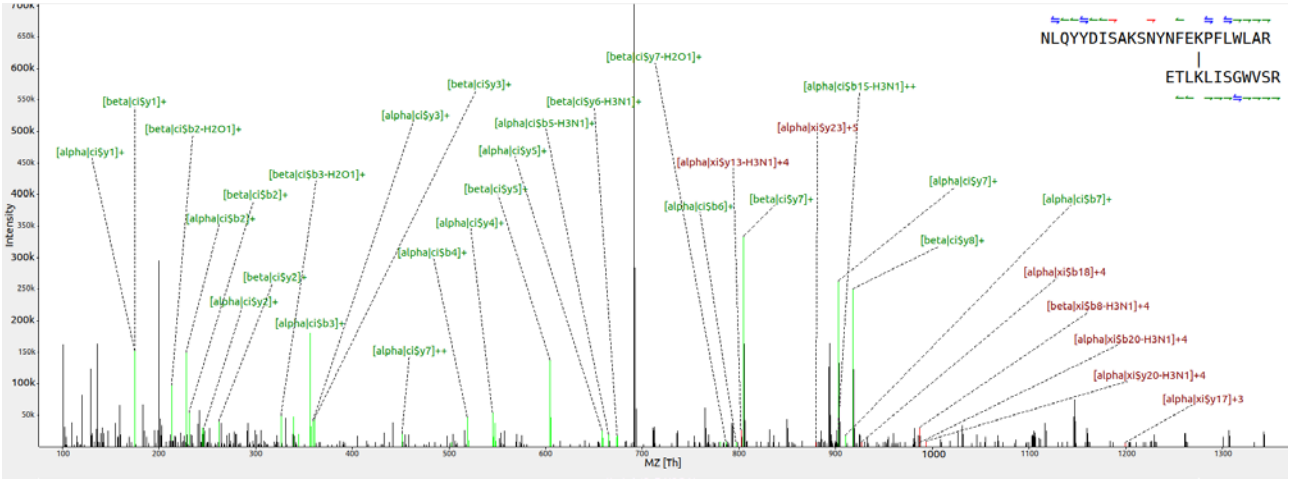
Supplemental Figure S10: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This URP was structurally validated.



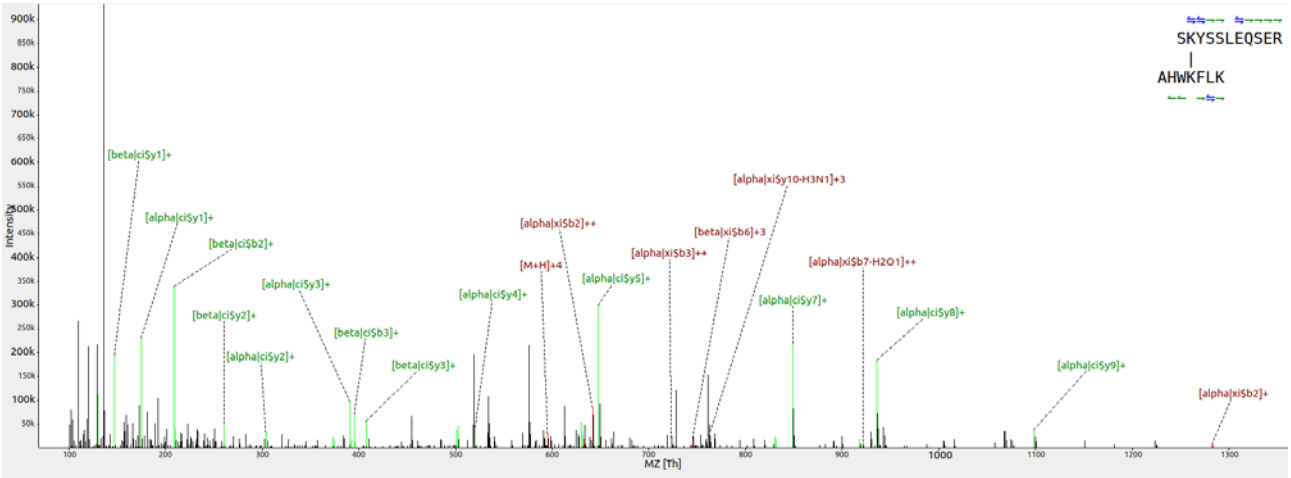
Supplemental Figure S11: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This URP was structurally validated.



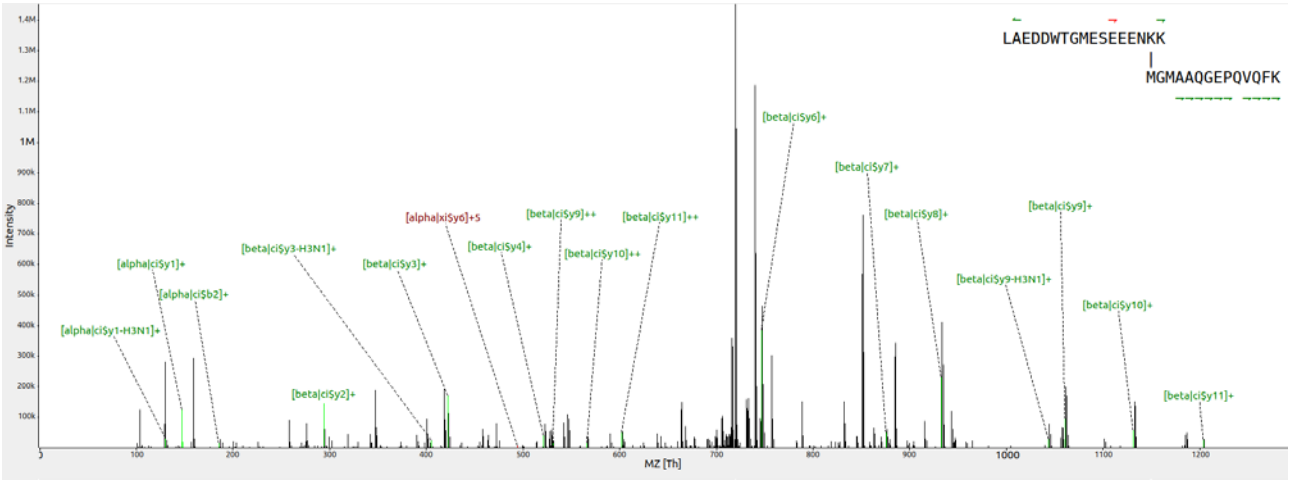
Supplemental Figure S12: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This URP was structurally validated.



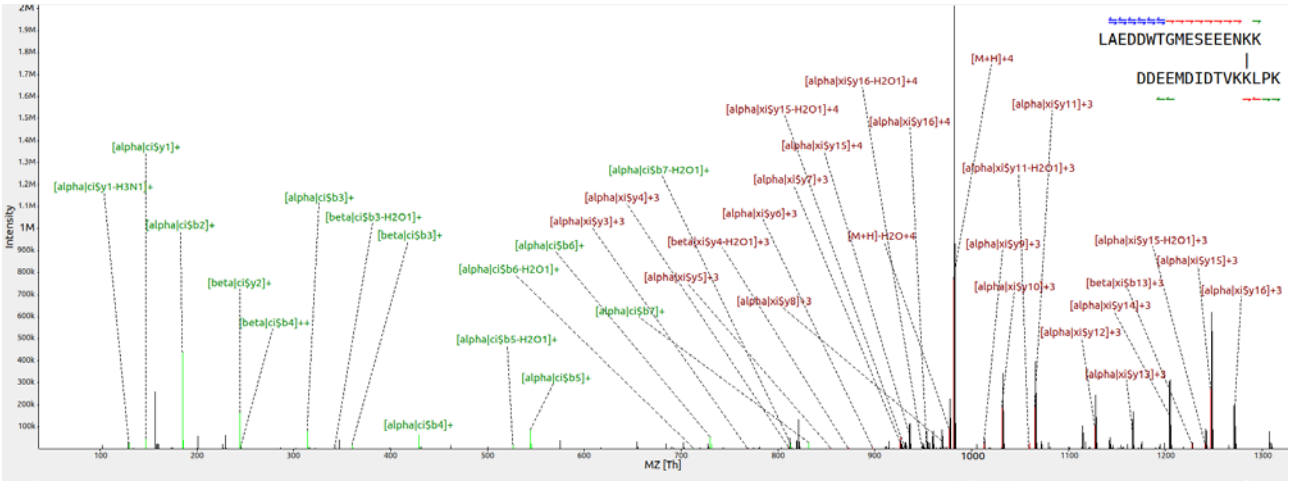
Supplemental Figure S13: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This URP was structurally validated.



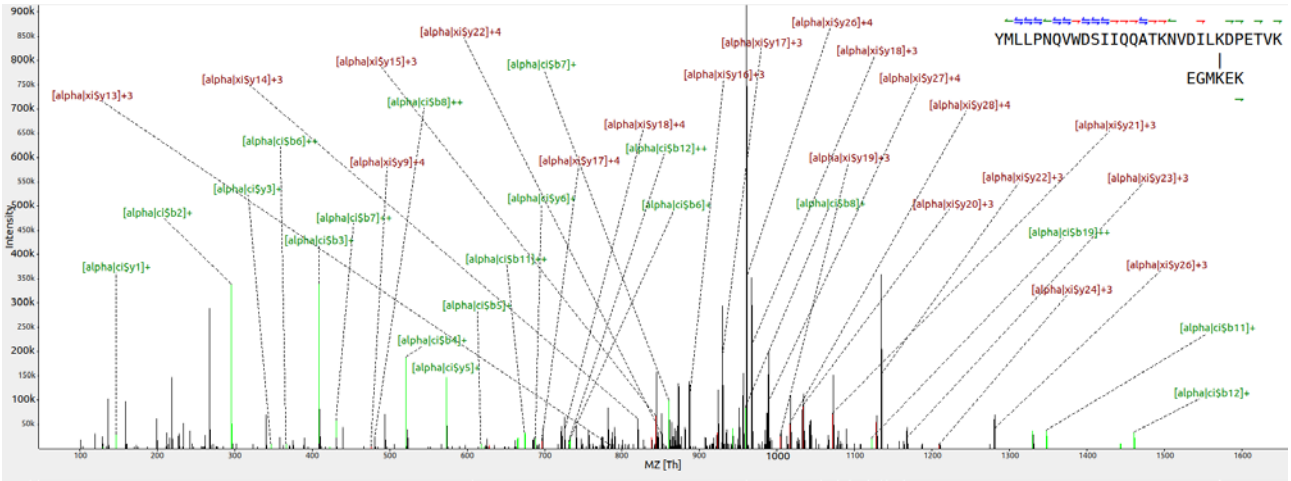
Supplemental Figure S14: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This URP was structurally validated.



Supplemental Figure S15: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.

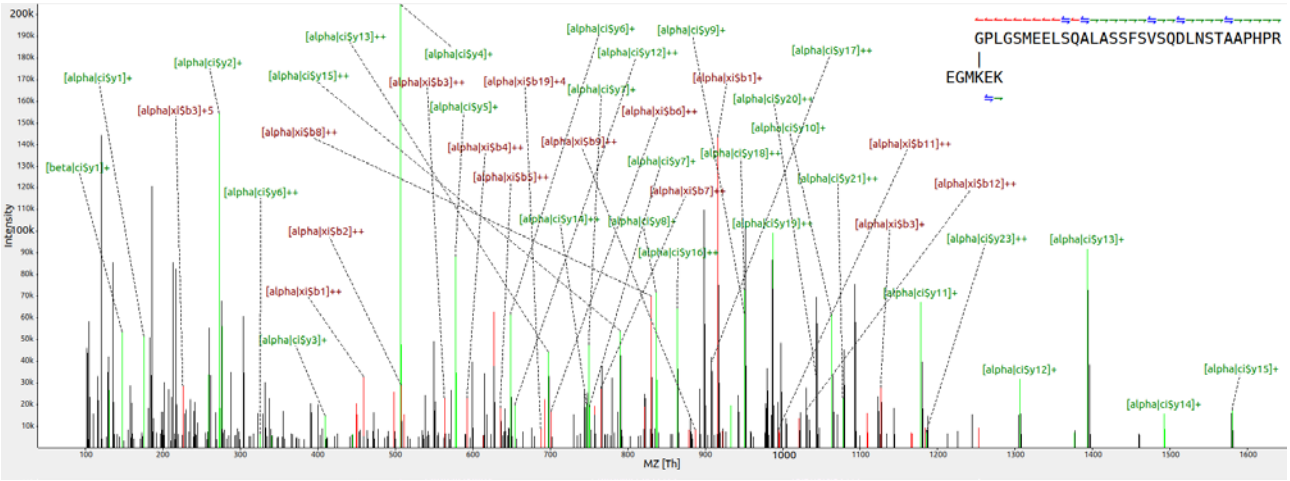


Supplemental Figure S16: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.

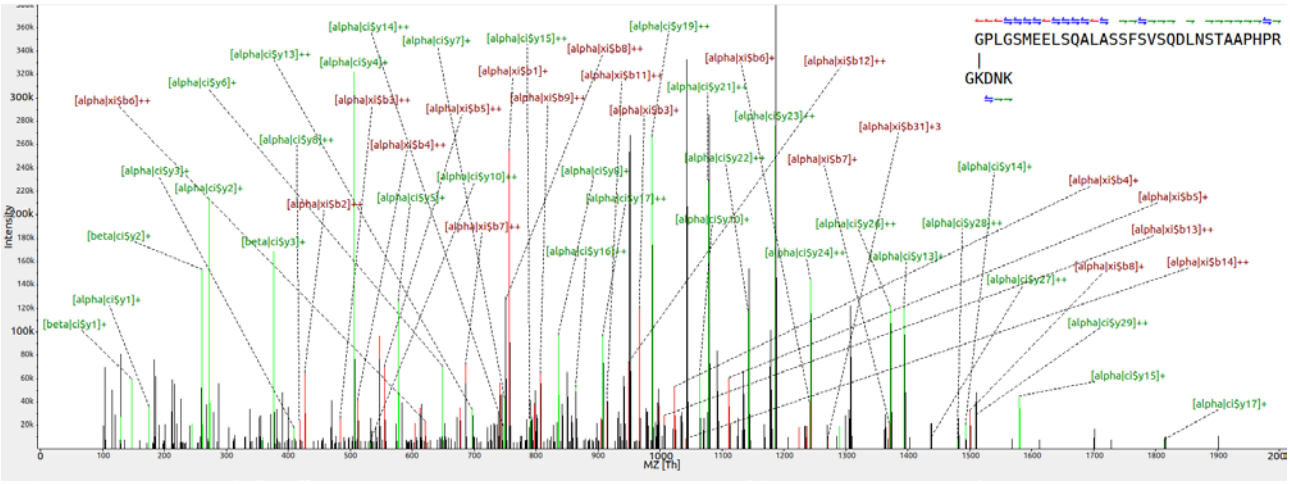


Supplemental Figure S17: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not

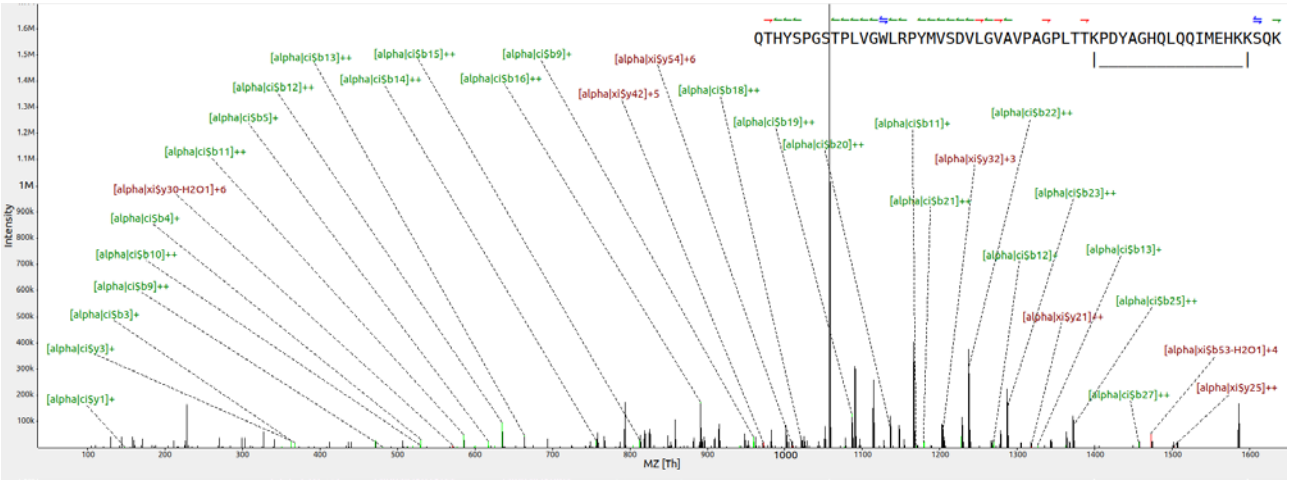
be validated.



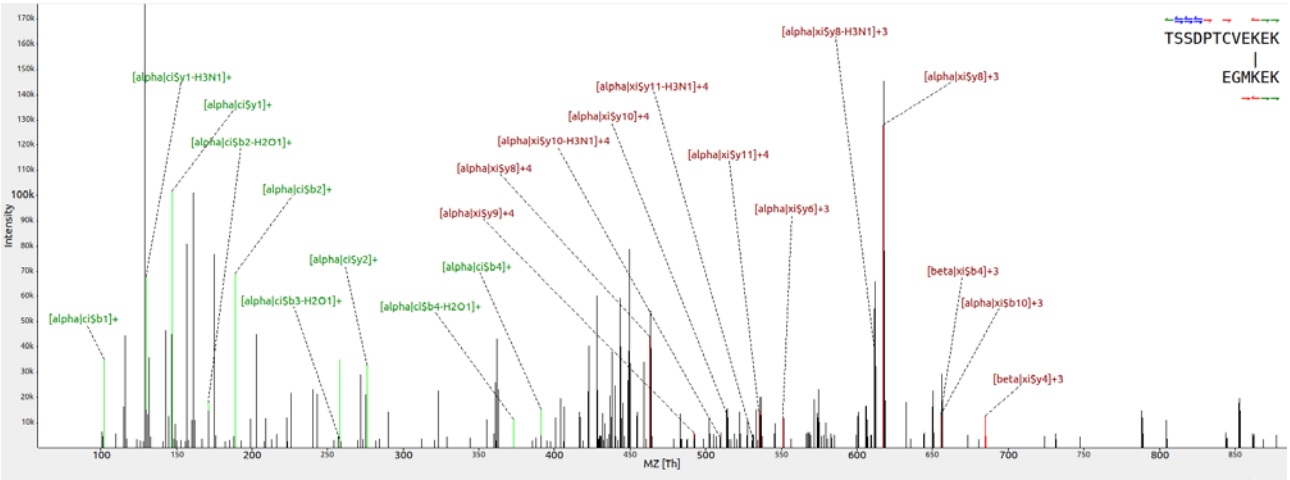
Supplemental Figure S18: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.



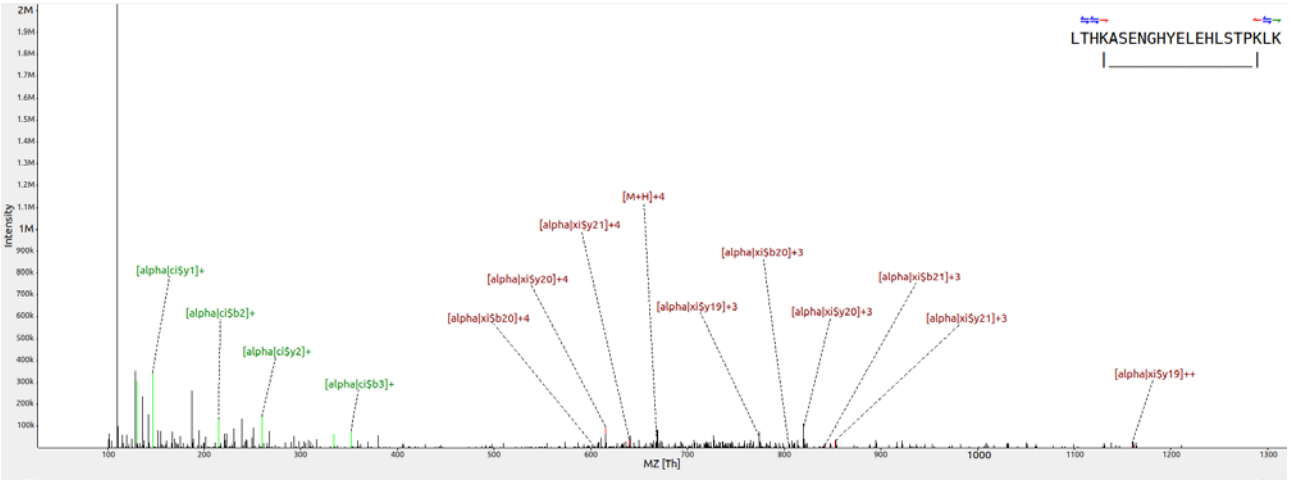
Supplemental Figure S19: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.



Supplemental Figure S20: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.



Supplemental Figure S21: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.



Supplemental Figure S22: The highest scoring CSM to one of the 14 URPs that were identified uniquely by OpenPepXL in the CRM dataset. This cross-link was not covered by the PDB structure and could therefore not be validated.

Supplemental Tables

Supplemental Table S1: Search parameters for the fractionated cell lysate dataset. The symbol '«' means this parameter for this tool was set to the same value as the first column.

	OpenPepXL 1.1	Kojak 1.6.0	pLink 2.3.5	XiSearch 1.6.731
precursor mass tolerance (ppm)	6	«	«	«
precursor charges	+3 to +8	NA	NA	NA
precursor monoisotopic peak corrections	0, 1, 2	0, 1, 2	(default, as this is not exposed as a parameter)	(default, as this is not exposed as a parameter)
fragment mass tolerance (ppm)	20	«	«	«
fixed modifications	Carbamidomethyl (C)	«	«	«
variable modifications	Oxidation (M)	«	«	«
max variable mods per peptide	3	«	«	«
enzyme	Trypsin	«	«	«
min peptide length	5	NA	5	5
max peptide length	NA	NA	600 (limited by digestion settings)	NA
max missed cleavages	4	«	«	«
min peptide mass	NA	300	300	NA
max peptide mass	NA	50000	50000	NA
cross-linker name	BS3	«	«	«
cross-linker mass	138.0680	«	«	«
mono-link masses	156.0786, 155.0946	156.0786	156.079	156.0786, 155.0946
linked residues	K, N-term	«	«	«

Supplemental Table S2: Search parameters for the BSA dataset. The symbol '«' means this parameter for this tool was set to the same value as in the first column.

	OpenPepXL 1.1	OpenPepXL 1.1	xQuest 2.1.3	xQuest 2.1.3	xQuest 2.1.3	xQuest 2.1.3
MS2 type	orbitrap	orbitrap	ion trap	orbitrap	ion trap	orbitrap
precursor mass tolerance (ppm)	10	«	«	«	«	«
precursor charges	+3 to +8	«	«	«	«	«
precursor monoisotopic peak corrections	0,1,2,3,4,5	0,1,2,3,4,5	NA	NA	NA	NA
fragment mass tolerance	20 ppm	20 ppm	0.2 Da	20 ppm	0.2 Da	20 ppm
cross-linked fragment mass tol.	20 ppm	20 ppm	0.3 Da	20 ppm	0.3 Da	20 ppm
fixed modifications	Carbamidomethyl (C)	«	«	«	«	«
variable modifications	Oxidation (M)	«	«	«	«	«
max variable mods per peptide	2	«	«	«	«	«
enzyme	Trypsin	«	«	«	«	«
min peptide length	5	«	«	«	«	«
max missed cleavages	2	«	«	«	«	«
cross-linker name	DSS	PDH	DSS	DSS	PDH	PDH
cross-linker mass	138.0680	152.1061	138.0680	138.0680	152.1061	152.1061
mono-link masses	156.0786, 155.0946	170.1167	156.0786, 155.0946	156.0786, 155.0946	170.1167	170.1167
linked residues	K, S, T, Y, N-term	D, E, C-term	K, S, T, Y, N-term	K, S, T, Y, N-term	D, E	D, E
isotopeshift	12.0753	10.0627	12.0753	12.0753	10.0627	10.0627
ntermxlinkable	NA	NA	1	1	0	0
Isopair_Mr_tolerance	NA	NA	15 ppm	15 ppm	15 ppm	15 ppm
Isopair_Tr_tolerance	NA	NA	3	3	3	3

Supplemental Table S3: Search parameters for the CRM dataset. The symbol '«' means this parameter for this tool was set to the same value as in the first column.

	OpenPepXL 1.1	Kojak 1.6.0	pLink 2.3.5	XiSearch 1.6.731	StavroX 3.6.6.5	xQuest 2.1.3
precursor mass tolerance (ppm)	10	«	«	«	«	«
precursor charges	+3 to +8	NA	NA	NA	NA	+3 to +8
precursor monoisotopic peak corrections	0,1,2	0,1,2	NA	NA	NA	NA
fragment mass tolerance (ppm)	20	«	«	«	«	«
fixed modifications	Carbamidomethyl (C)	«	«	«	«	«
variable modifications	«	«	«	«	«	«
max variable mods per peptide	2	«	«	«	«	«
enzyme	Trypsin	«	«	«	«	«
min peptide length	5	NA	5	5	5	5
max peptide length	NA	NA	300	NA	NA	NA
max missed cleavages	2	«	«	«	«	«
min peptide mass	NA	300	300	NA	300	NA
max peptide mass	NA	50000	50000	NA	50000	NA
cross-linker name	BS3	«	«	«	«	«
cross-linker mass	138.0680	«	«	«	«	«
mono-link masses	156.0786, 155.0946	156.0786	156.079	156.0786, 155.0946	156.0786	156.0786, 155.0946
linked residues	K, N-term	«	«	«	«	«
consecutive peptides	NA	NA	NA	NA	1	NA
Isopair_Mr_tolerance	NA	NA	NA	NA	NA	0
Isopair_Tr_tolerance	NA	NA	NA	NA	NA	0.02
isotopeshift	NA	NA	NA	NA	NA	0
printisotopicscanpairs	NA	NA	NA	NA	NA	0
printlightonlypairs	NA	NA	NA	NA	NA	1
ntermxlinkable	NA	NA	NA	NA	NA	1

Supplemental Table S4: Number of CSMs identified in the synthetic peptide dataset at a 5% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al*[3]. Kojak was omitted, because CSM level results for Kojak were not available in that publication.

Search engine	Number of cross-links								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	822	938	895	28	24	34	3.3	2.5	3.7
pLink2	639	712	683	27	27	39	4.1	3.7	5.4
StavroX	378	434	419	9	12	10	2.3	2.7	2.3
XiSearch	491	498	547	20	13	10	3.9	2.6	1.8

Supplemental Table S5: Number of URPs identified in the synthetic peptide dataset at a 5% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al*[3].

Search engine	Number of cross-links								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	242	250	237	21	21	21	8.0	7.7	8.1
pLink2	217	230	203	26	24	33	10.7	9.4	14.0
Kojak	220	225	217	68	60	67	23.6	21.0	23.6
StavroX	159	175	154	8	10	9	4.8	5.4	5.5
Xi	179	183	179	18	11	7	9.1	5.7	3.8

Supplemental Table S6: Number of CSMs identified in the synthetic peptide dataset at a 1% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al*[3]. Kojak was omitted, because 1% FDR results for Kojak were not available in that publication.

Search engine	Number of cross-links								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	368	506	365	4	5	4	1.1	1.0	1.1
pLink2	594	644	585	10	13	25	1.7	2.0	4.1
StavroX	265	157	160	4	0	1	1.5	0	0.6
Xi	312	352	438	2	4	5	0.6	1.1	1.1

Supplemental Table S7: Number of URPs identified in the synthetic peptide dataset at a 1% FDR cut-off. All data except for OpenPepXL was taken from Beveridge *et al*[3]. Kojak was omitted, because 1% FDR results for Kojak were not available in that publication.

Search engine	Number of cross-links								
	Correct			Incorrect			Calculated FDR (%)		
	R1	R2	R3	R1	R2	R3	R1	R2	R3
OpenPepXL	161	196	148	2	4	3	1.2	2.0	2.0
pLink2	215	218	189	9	12	22	4.0	5.2	11.6
StavroX	124	91	90	4	0	1	3.1	0	1.1
Xi	141	152	163	2	3	5	1.4	1.9	3.0

Supplemental Table S8: Targets and decoys assigned to spectra in total, as well as validated CSMs above and Unique Residue Pairs assigned below the 5% FDR cut-off for the first replicate (R1) of the synthetic peptide dataset. All data except for OpenPepXL was taken from Beveridge *et al*[3]. The number of assigned decoys for StavroX is missing, because they are not reported by StavroX.

Search engine	Total target CSMs	Total decoy CSMs	Validated CSMs above 5%	Additional URPs below 5%
OpenPepXL	2029	2156	822	80
pLink2	1006	383	639	4
StavroX	1322	n.a.	378	58
Xi	1686	2677	491	11

References

1. Monecke, Thomas, *et al.* "Crystal structure of the nuclear export receptor CRM1 in complex with Snurportin1 and RanGTP." *Science* 324.5930 (2009): 1087-1091.
2. Walzthoeni, Thomas, *et al.* "False discovery rate estimation for cross-linked peptides identified by mass spectrometry." *Nature Methods* 9.9 (2012): 901.
3. Beveridge, Rebecca, *et al.* "A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes." *Nature communications* 11.1 (2020): 1-9