Supporting Information: In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides

Swantje Lenz[1], Sven H. Giese[1], Lutz Fischer[2], and Juri Rappsilber*[1, 2]

[1]Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany

[2]Wellcome Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

List of Supporting Information

*juri.rappsilber@tu-berlin.de

MS1 based mass range reduction

Since considering multiple precursor masses increases search time, we developed an approach to reduce the range searched in Xi-MPA. For each MS2 spectrum, the precursor peak is identified in the corresponding MS1 spectrum. Then, the most abundant occurrence of this peak is searched in a retention time window of 20 seconds (±10 seconds) and the corresponding MS1 spectrum is extracted. Assuming that the assigned charge state was correct, the newly extracted MS1 spectrum is searched for the true monoisotopic peak, i.e. lighter isotope peaks than the one that was reported in the MS2 spectrum. According to Table S4 each MS2 spectrum gets assigned an individual range of possible precursor masses to be used during Xi-MPA.

Table S4: Xi-MPA mass range reduction.

Lighter Peaks Present	**Search Range**
none	mass range without lightest mass*
continuous (without gaps)	lightest two peaks found
single peaks (with gaps)	mass range up to lightest peak found

* In case the range this approach is done is only up to -2 Da, -1 Da will still be searched here.

This approach was evaluated on the first and last fractions of dataset 3 with a mass range of up to -4 Da. On average, the masses searched in Xi-MPA were reduced by 24% per file, while the number of within PSMs is 97% of the search without range reduction.

Note that this approach increases the time of the preprocessing before search to some extent. Therefore, it is only worthwhile undertaking for searches with a large database, for which the time of the search itself is long.

In previous approaches we tried to incorporate a mass and / or intensity cutoff or dependency when selecting the considered mass range. However, the applied heuristics resulted in significant losses in PSM numbers, presumably because a clear cut in the distributions is missing (Figure 4A and S5).

Python scripts were written using the pyopenms package [1] and are available under
https://github.com/Rappsilber-Laboratory/Xi-MPA_scripts.

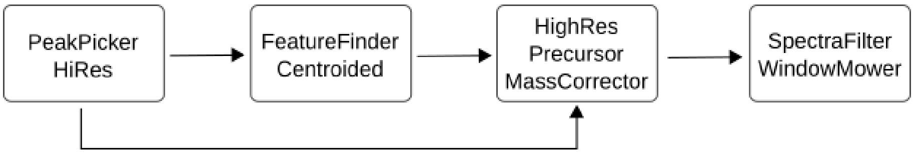

Figure S1: OpenMS preprocessing workflow. PeakPickerHiRes was used with the following settings: 'ms levels' was set to 1 and 'signal to noise' set to 0 (disabled). For the tool FeatureFinderCentroided the following settings were changed: 'feature:min score' - 0.6, 'mass trace:min spectra' - 7, 'isotopic pattern:charge low' - 3, and 'isotopic pattern:charge high' - 7. In HighResPrecursorMassCorrector 'feature:rt tolerance' was changed to 15. SpectraFilterWindowMower was used with 'movetype' - jump, 'windowsize' - 100, and 'peakcount' 20.

test statistic	p-value	H_1	Data	significant
35196.5	2.54E-29	less	0 vs. ref	True
36626.5	4.68E-27	less	-1 vs. ref	True
38792.5	6.30E-32	less	-2 vs. ref	True
39451.0	4.66E-25	less	-3 vs. ref	True
28239.5	4.78E-19	less	-4 vs. ref	True

Table S5: Summary for conducted significance tests for Fig. 3B in the main text. The significance level α was set to 0.05 before the statistical analysis. The Wilcoxon rank sum test with continuity correction was used in R. Abbreviations: ref - reference distance distribution derived from all cross-linkable residues. 0, -1, -2, -3 and -4 denote the subsets of PSMs with the corresponding mass shift of the precursor.

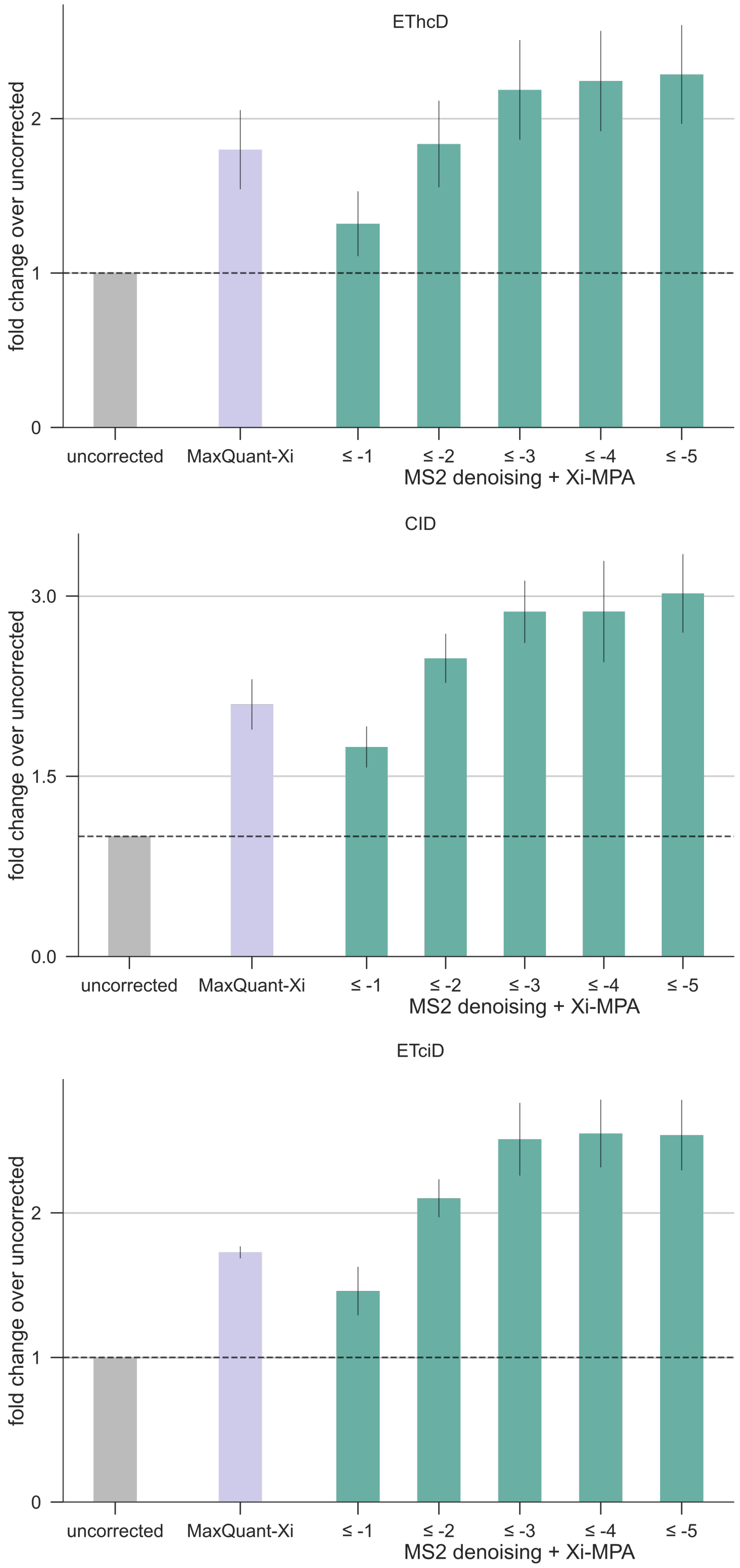

Figure S2: Performance of Xi-MPA on EThcd, CID, and ETciD acquired data of the pseudo-complex dataset. Different ranges in Xi-MPA were tested and evaluated on the number of PSMs. Shown is the mean fold change of the respective setting to the number of PSMs from unprocessed data. For all fragmentation methods, the number of identifications increases compared to the unprocessed data. While for EThcD 251 PSMs were identified for the unprocessed data, 434 PSMs resulted from MaxQuant-Xi and 542 PSMs from Xi-MPA with up to -4 Da. Numbers of identified PSMs for CID data are: 265 PSMs for unprocessed, 552 for MaxQuant-Xi, and 753 for Xi-MPA (-4 Da). Finally, 197 PSMs are identified in unprocessed data for ETciD, while 340 resulted from MaxQuant-Xi and 502 from Xi-MPA. Although the increase of Xi-MPA is smaller for EThcD and ETciD data than for CID and HCD data, it is the approach with the most identifications.

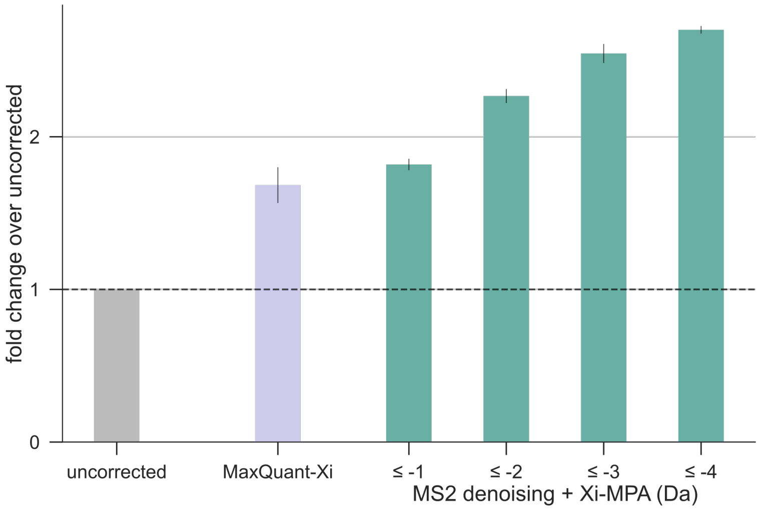

Figure S3: Performance of Xi-MPA on HCD acquisitions of the HSA dataset. The dashed line equals a fold change of 1, meaning the same number of PSMs as in the unprocessed data was identified. Different ranges of Xi-MPA were tested and compared to MaxQuant-Xi results. While the latter led to 1127 PSMs, Xi-MPA with up to -4 Da resulted in 1816 identifications.

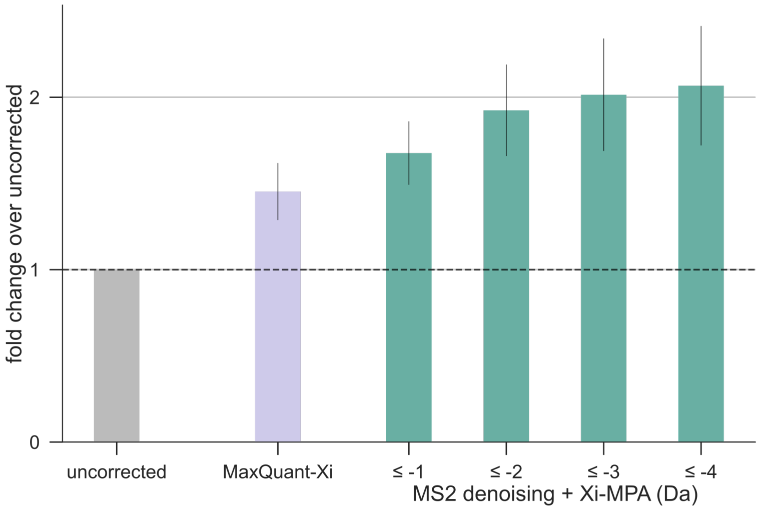

Figure S4: Performance of Xi-MPA on the first and last fraction of the *C. thermophilum* dataset. As for the other two datasets, different ranges of Xi-MPA were compared to MaxQuant-Xi results. MaxQuant led to 2966 identifications, while Xi-MPA with -4 Da identified 4013 PSMs. Considering masses up to -3 Da led to a similar number of PSMs (3945) than up to -4 Da.

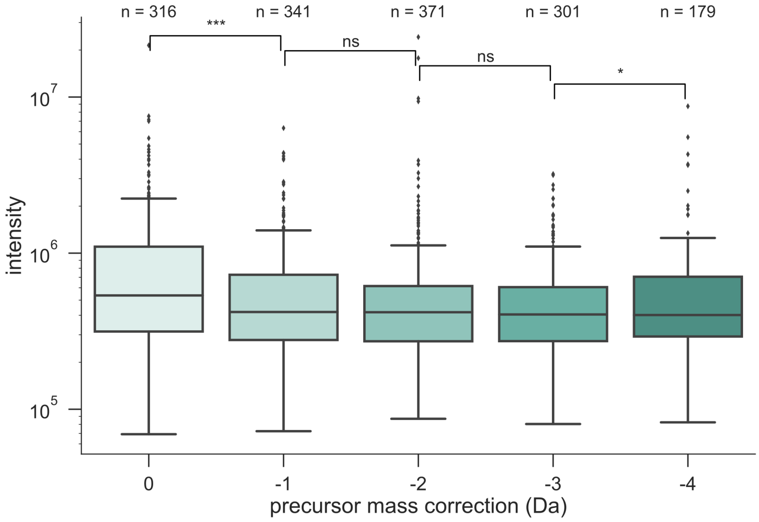

Figure S5: Dependency of the monoisotopic mass correction on precursor intensity. Scans of the pseudo-complex dataset identified in the Xi-MPA search were evaluated regarding their mass correction. Corrections to lighter masses occur more often for precursors with lower intensity. Significance is denoted by asterisks (ns: p-value>0.05, *: p-value<0.05, ***: p-value<0.001).

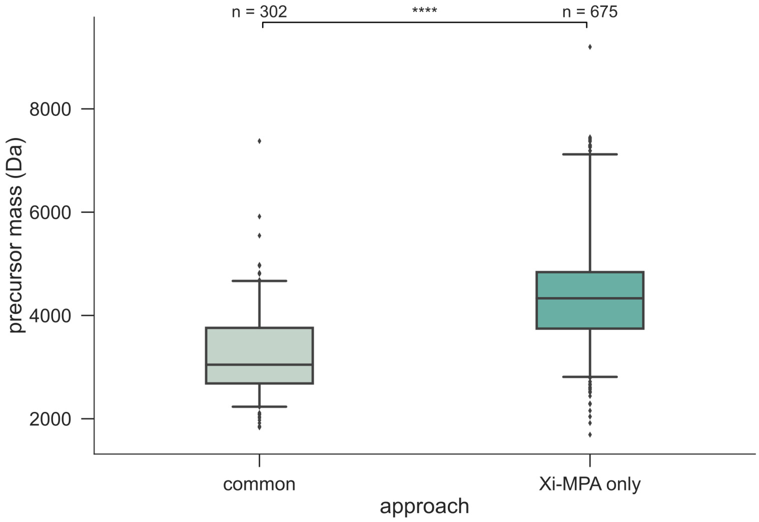

Figure S6: Correction of the monoisotopic mass is more successful for lighter peptides, while Xi-MPA identifies larger peptides more often. Precursor masses of scans identified in all approaches (preprocessing in MaxQuant and OpenMS and Xi-MPA) are compared to scans solely identified in Xi-MPA. (****: p-value<0.0001)

References

[1] Hannes L. Röst, Uwe Schmitt, Ruedi Aebersold, and Lars Malmström. pyopenms: A python-based interface to the openms mass-spectrometry algorithm library. *PROTEOMICS*, 14(1):74–77, 2014.