**Computational fragmentation spectra prediction for identifying the unknowns from LC-high resolution MS data**

Meng Hu1,2, Tobias Schulze1, Emma Schymanski3, Christoph Ruttkies4, Erik Müller1, Steffen Neumann4, Werner Brack1, Martin Krauss1

1 Helmholtz Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany

2 RWTH Aachen University, Institute for Environmental Research, Worringerweg 1, 52074 Aachen, Germany

3 Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland

4 Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle(Saale), Germany

**Key words:** LC-HRMS, fragmentation prediction, environmental contaminants

Despite recent advances in high resolution mass spectrometry coupled with liquid chromatography (LC-HRMS), the identification of small molecules from tandem MS data is still seen as a bottleneck. Structure retrieval based on molecular formula from comprehensive compound databases leads to numerous candidate structures. Hence, a selection of the most probable or unlikely candidate structure(s) by tandem fragmentation prediction is nessesary. Both commercial and free tools are capable of this purpose, but their performance for environmental contaminants has hardly been assessed so far. In this study, we applied and compared two software-based fragmentation prediction tools to evaluate their performance.

We analyzed a dataset containing 78 known micropullutants with widely varying structures, for which MS/MS spectra were measured in ESI+ and ESI- mode on a Thermo Scientific LTQ Orbitrap XL instrument using different fragmentation techniques (CID and HCD). Candidate structures for each compound were retrieved from the ChemSpider database and saved as sdf format file. The sdf file and tandem MS spectra of the same compound were imported into MetFrag and CFM, which are two combinatorial fragmenters to enumerate all possible fragments of the structure by systematically breaking bonds and ranking the candidates according to the agreement between predicted and experimental mass spectra. The score and rank of each candidate and the “correct” candidate structure were extracted by a R script written for this purpose. The results show that both MetFrag and CFM have good score for about 80% of the compounds, however for several compounds with neither software a good score was obtained. Generally, MetFrag performs slightly better than CFM for the ESI+ dataset and significantly for the ESI- dataset in terms of score and the relative rank of the correct candidate structure. One possible explanation for this finding is that the training dataset of current version CFM was acquired on a QToF instrument in positive mode. This indicates that the performance of CFM might be improved if it was trained on both positive and negative mode Orbitrap micropollutant data for this case. Furthermore, the combination of MetFrag and CFM increased the rank of correct candidate structures for around 25% of the compounds, while it decreased the rank of the correct structure for about 10% of the compounds. In addition, we found that merged mass spectra from different collision energies did not improve the rank and score significantly. For negative ion mode MS dataset, both softwares performed much better for HCD fragmentation than for CID in most cases, probably due to the fact that more fragments were generated by HCD than by CID.