Finding small molecules in big data – smarter

Emma L. Schymanski[1] & Antony J. Williams[2]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 7 avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg. emma.schymanski@uni.lu
[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711 USA

Metabolomics and exposomics are amongst the youngest and most dynamic of the omics disciplines. While the molecules involved are smaller than proteomics and the other, larger "omics", the challenges are in many ways greater. Elements are less constrained, there are no given "puzzle pieces" and there is a resulting explosion in terms of potential chemical space. It is impossible to even enumerate all chemically possible small molecules. The challenges and complexity of identifying small molecules even using the most advanced analytical technologies available today is immense. Current "big data" methods for small molecules rely heavily on chemical databases, the largest of which presently available contain ~100 million chemicals. Despite this large number, high resolution mass spectrometry (HR-MS) measurements contain tens of thousands of features, of which only a few percent can be annotated as "known" and confirmed as metabolites or chemicals of interest using these chemical databases. How can we find relevant small molecules in the ever increasing data loads? How can we annotate more of the unknown features in HR-MS experiments? This talk will present European, US and worldwide initiatives to help find small molecules in big data in smarter ways - from chemical databases to spectral libraries, real-time monitoring to retrospective screening. It will touch on the challenges of standardized structure representations, data curation and deposition. Finally, it will show how interdisciplinary communication, data sharing and pushing the boundaries of current capabilities can facilitate research efforts in metabolomics, exposomics and beyond. *This abstract does not necessarily represent U.S. EPA policy*.