Review Kim et al., AMTD 2022

“Comprehensive detection of analytes in large chromatographic datasets by coupling factor analysis with a decision tree”

Kim et al. combined positive matrix factorization (PMF) and a decision tree for comprehensive peak detection in GC-MS datasets. Therefore, chromatograms are sliced in short sections and within these sections a certain (variable) number of factors is predetermined. These factors then contain a chromatographic profile along with its (fragmentation) mass spectrum. A decision tree algorithm discards factors that represent no compound. The combined PMF/decision tree data evaluation tool is successfully tested on a chromatogram of poorly separated deuterated tetradecane as well as on a complex ambient GC-MS dataset of the GreenOcean-Amazon field campaign.

Overall, the paper is well written, precise and comprehendible, even for non-PMF experts. The quality of the graphs is good and I can recommend the paper being published in AMT after addressing the following minor comments:

Minor comments:

l. 17: Why is “peak width” not included in peak evaluation of the decision tree?

l. 23 & l. 370: It is mentioned that 90% of the ~1100 “analytes” have no match with the NIST-database. Therefore, I suggest to use the term “features” instead of “analytes”. This applies to the whole manuscript.

l. 30-35 or somewhere else in the introduction: Might be worth mentioning 2D-GC approaches to resolve complex samples.

l. 170 ff.: Peak-shapes of chromatograms are important and mostly not a perfect Gaussian. Peak shapes depend on the properties of the compound, but also on the condition of the column. That means, when analysing a few hundred of samples, the peak shape for one compound can become worse over time. How does the algorithm deal with that?
The authors used a Gaussian for chromatographic peak fitting and mention that a more complex approach is not necessary, although possible to include modified peak shapes. I disagree with the statement that non-Gaussian peak models are not necessary for proper peak fitting. An exponentially modified Gaussian (EMG) actually allows evaluating the Gaussian shape of chromatographic peaks by fitting with four variables (area, elution time, peak width and exponential) instead of just area, elution time and peak width (e.g. see Goodman & Brenna 1994). The mentioned paper by Isaacman-Van Wertz et al. (2017) is missing in the references.
How robust is the finding of the three different compounds ($C_{14}D_{30}$, $C_{14}D_{29}H$, $C_{14}D_{28}H_2$) of the nine-factor solution when a non-Gaussian peak shape model is employed that allows to fit peak tailing? Since the difference is in the mass spectrum at *m/z* 226-230, I assume it is robust, but I can be wrong.

l. 174: The authors should provide evidence that “a refined peak shape is likely unnecessary”, or otherwise should argue more carefully.

l.227: “M1 and M2 are normalized mass spectra of two analytes”. This is confusing, because if the result is that epsilon>=0.8, then it is two normalized mass spectra of one analyte. I suggest rephrasing “M1 and M2 are normalized mass spectra selected for comparison”.

Figure 5 shows several low-abundant compounds that were detected manually (blue asterisks). Were these compounds also detected by the presented PMF method? Why are the large prominent peaks in the TIC not detected by the manual method? It looks if there is a homologue series of alkanes in the chromatogram (visible as an evenly-spaced series of decreasing peaks from 400-650 s). Why has this not been identified in the manual analysis? As a consequence, the manual inspection could easily identify much more compounds, with implication on the statement of the “one order of magnitude” (line 381).

Technical notes:

l. 30: HüBschmann → Hübschmann.

l. 249-250: use a non-breaking space between number and unit.

Literature:

Goodman KJ, Brenna JT (1994) Curve fitting for restoration of accuracy for overlapping peaks in gas chromatography/combustion isotope ratio mass spectrometry. Anal. Chem. 66(8):1294–1301.