Going deeper in tile-based processing for complex GC×GC-TOFMS datasets

Meriem Gaida, Caitlin N Cain, Robert E Synovec, Jean-François Focant, **Pierre-Hugues Stefanuto**

In the quest of making multidimensional chromatography (MDC) a robust method for untargeted screening of small molecules, one of the key remaining challenges to tackle is reproducibility. To reach this objective, important analytical aspects, such as column dimension and separation conditions need to be investigated. The biggest challenge for MDC is nevertheless data processing, meaning transforming row data into pertinent information. To enable data analytical method and processing workflow evaluation, a reference data set is required. In this study, we used a whole stool research grade test materials (RGTMs) prepared by NIST for interlaboratory studies to develop a control data set covering sampling, analysis, and processing workflows. The RGTMs contain two diets, vegan and omnivore, and two sample formats, liquid vs lyophilized. In this presentation, we will focus on the utilization of data produced from RGTMs to evaluate data processing approaches.

The robustness of several statistical workflows involving commercial, in house, and open-source solutions were investigated. First, we investigated user impact on a well-established ANOVA-based workflow. The key was to evaluate the weight of human decision on the final classification metrics and the impact on the identification of significant features. Our well-established workflow shown to be unimpacted by human decision during data cleaning, pre-processing and model building as no significant output changes appeared in the study.

Next, we developed and evaluated a new processing approach combining tile-based image comparison and machine learning-based feature selection. The combination of tile-based alignment and random forest classification increased the robustness, compared to the ANOVA-based approach. Indeed, the false positive rate decreased during feature selection, and we were able to conduct unbalanced data set processing.