Unraveling chemical structure-property relationships from two-dimensional liquid chromatography and mass spectrometry (HIMS)

Intended promotor: *dr. Bernd Ensing (HIMS); dr.Joris Mooij (IvI)*
Intended co-promotor: *dr. Bob Pirok (HIMS); lab manager*

Background:

Two-dimensional liquid chromatography (2DLC) is a powerful technique to separate and detect trace molecular compounds in complex samples. The separation is based on the difference in "retention" time that it takes for each compound to be carried by a solvent through a column filled with a material that interacts with the injected compounds based on a chosen property, such as molecular size or hydrophilicity.

In the HIMS Analytical Chemistry Group, 2DLC is combined with mass spectroscopy (MS) to detect molecules in highly complex samples, e.g. from food, polymers or pharmaceuticals that contain thousands of different unknown compounds. As the compounds can have extremely low concentrations (e.g. protein biomarkers, plant hormones, food contaminants), the signals of interest are often buried in the noise of the data and information is being lost. Typical datasets contain several gigabytes of data per measurement.

To analyse such data and to extract all relevant information, new techniques are required. One complicating factor is that successful implementation of the technique requires months of costly and cumbersome development. In response to this, algorithms are being developed to model chromatographic interaction of analyte molecules with the chemical moieties of the stationary phase, so as to allow prediction of optimal chromatographic conditions. Successful modelling of chromatographic retention times requires data-processing algorithms to locate analyte molecules across several data sets.

Can the analysis of 2DLC/MS data to recognize peaks, components, and even artifacts be automatized to parameterize the 2DLC for optimal resolution? We propose to develop and apply AI machinery to unravel structure-property relationships from these data sets that can be used to predict retention times and generate 2DLC optimization schemes.

Aim:

Develop an optimization scheme for 2DLC separation and peak refinement as a function of e.g. the choice of stationary materials, temperature, pressure, (non-linear) time-dependent solvent mixture composition, and solvent flux.

Approach:

We approach our aim, the development of an optimal 2DLC separation strategy (or "method"), in four work packages (WP). To control the feasibility at each stage, we start with 1DLC before embarking on the 2DLC problem, and with relatively simple samples (e.g. dye extracts) that can then be made increasingly complex and challenging to separate (e.g. metabolite separation or peptide characterizations). Initially, the PhD student will join analytical scientists in the lab to learn the concepts of the applied 2DLC technology. The deliverables of each WP are (1) open-source software that can be made publicly available, and (2) publication in a scientific journal that can be either in the field of analytical chemistry or of computer science.

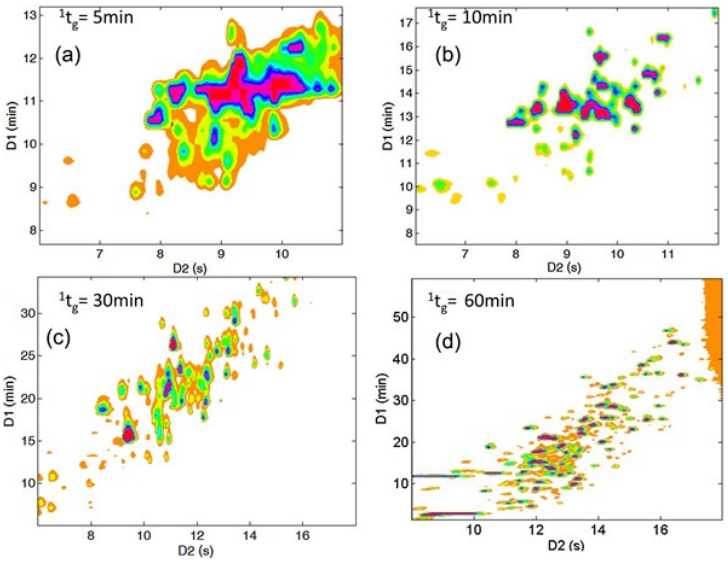

Figure 2: Separation of a peptide digest sample by 2DLC using four different gradient times (tg) in the first dimension. The gradient time determines the length over which a solvent-composition program is carried out and is just one out of a large number of method parameters, to maximize the peak resolution.

WP 1. Develop a machine learning tool for rapid peak identification in (noisy) retention time data, recognizing artifacts (e.g. from breakthrough, leading to spurious splitting of peaks), and estimating the relative amounts of species in the samples. MS is used to label the data.

WP 2. Predict retention times of sample components as a function of 2DLC parameters, such as static material, solvent composition, flow rate, etc. Extract physical relationships between the molecular structure and the physico-chemical properties.

WP 3. Optimization of 2DLC parameters for maximum peak separation and minimal experimental run times.

WP 4. Analyze 2DLC data labeled by ultraviolet-visible (UV/vis) light spectra (instead of MS), which contains less clear information, but is used more in practice.

Impact:

If successful, the impact of the proposed innovations will be profuse. Retention time prediction and 2DLC process control will be a huge improvement on the current months of labour-intensive tuning for optimal peak capacity, which currently prohibits full 2DLC

exploitation for commercial applications. On the more fundamental side, analysis of the trained AI systems can uncover novel structure-property relations that go beyond the current empirical formulas used to predict retention times. Smart interpretation of 2DLC data sets allows for extraction of a wealth of information that with current limited methods is unfortunately disregarded. In addition, we expect that (parts of) the resulting AI systems will not be limited to the proposed 2DLC application, but can be relatively easily extended to other chemical, biological, and physical experimental data-sets.

Possibly interested external parties:

DSM, NFI & Shell

References:

[1] Sarrut et al. *J. Chromatogr. A*, 1498, **2017**, 183-195.