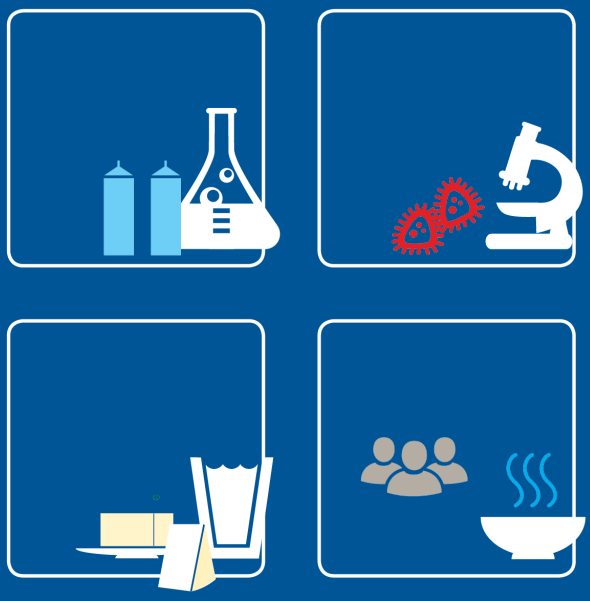
SLUTRAPPORT/FINAL REPORT
NR. 2023-175

Rasmus Bro:

AUROMA – Automatiseret analyse af aromaprofiler med machine learning

AUROMA – Automatic analysis of aroma profiles using machine learning



Mejeribrugets
ForskningsFond

Juli 2023

Final report

for collaborative projects funded via the Danish Dairy Research Foundation (DDRF)

1. Title of the project

Dansk: AUROMA – Automatiseret analyse af aromaprofiler med machine learning

English: AUROMA – Automatic analysis of aroma profiles using machine learning.

2. Project manager

Rasmus Bro
University of Copenhagen, Faculty of Science, Department of Food Science
Rolighedsvej 26, 1958 Frederiksberg

+ 4535333296
rb@food.ku.dk

3. Other project staff

Jesper Løve Hinrich (Postdoc), KU Department of Food Science
Dillen Augustijn (Postdoc), KU Department of Food Science
Peter Bæk Skou (Postdoc), KU Department of Food Science

4. Sources of funding

Arla Foods, Mælkeafgiftsfonden

5. Project period

Project period with DDRF funding:	March/2019 to March/2021
Revised, if necessary:	[Start, End, month/year]
Total project period, if sub-project within a larger project:	[Start, End, month/year]
Revised, if necessary:	[Start, End, month/year]

6. Project summary

Dansk:

I dette projekt har vi brugt kunstig intelligens kombineret med analytisk kemisk erfaring for a lave en mere automatiseret analyse af de data, man får fra en såkaldt gaskromatografisk analyse. Tidligere var sådan en analyse meget tidskrævende (dage), men vi har automatiseret dette i en grad, så der kun kræves maksimalt få timer. Ydermere er de data, man får ud af det langt, langt bedre end det, man fik før. Man får information om flere kemiske stoffer, og man får disse stoffer identificeret og kvantificeret langt bedre.
Vi har også fået proof of concept på, at vi kan automatisere analysen fuldstændig og dermed helt eliminere den menneskelige faktor.

English:

In this project, we have used artificial intelligence combined with our skills in analytical chemistry to make a more automated analysis of the data obtained from a so-called gas chromatographic analysis. In the past, such an analysis was very time-consuming (days), but we have automated this to a degree so that only a few hours are required at the most. Furthermore, the data you get out of it is far, far better than what you got before. You get information about more chemical substances, and you get a more precise identification and quantification of these substances.

We have also obtained proof of concept that we can completely automate the analysis and thus eliminate the human factor.

7. Project aim

Dansk:

Projektet omhandler automatisering af dataanalyse til kemisk fingerprinting af aromaprofiler ved brug af gaskromatografi (GC) med massespektrometrisk (MS) detektor. Projektet kombinerer matematik, kunstig intelligens, kemometri og softwareudvikling med analytisk kemi for at forbedre karakterisering af kemiske stoffer, så som kemiske stoffer der overlapper, med lav koncentration eller ved prøve uden kendte analytter. Samtidig skal dette ske på en måde, der er hurtig, reproducerbar, let tilgængelig og med minimal brug af analytikerens tid.

English:
The project concerns automation of data analysis for chemical finger printing of aroma profiles via gas chromatography (GC) coupled with a mass spectrometry (MS) detector. The project combines mathematics, artificial intelligence, chemometrics, and software development with analytical chemistry to improve characterization of chemical compounds, such as chemicals which co-elute, are present in low concentrations, or samples without known analytes. This has to be done in a timely and reproducible way that is easily accessible and minimizes the time spent by an expert analyst.

8. Background for the project

In this project, we will be using untargeted chemical profiling based on gas chromatography with mass spectrometric detection (GCMS). The data resulting from such instrumental systems have been shown to be inadequately analysed using traditional approaches. Severely overlapping chemical peaks are not resolved; low concentration analytes are not detected; samples void of analytes are incorrectly labelled etc. Moreover, the current approaches are also extremely time consuming and not robust to user influence. There is a need for more timely, reproducible, accessible and comprehensive methods for extracting the chemical information.

We will combine mathematics, statistics, chemometrics and computer science with analytical chemistry. Our approach will provide a new methodology for automatically analysing GCMS data.

Untargeted GCMS often gives complex data. No specific chemical analytes are in focus a priori and hence it is not possible to optimize the procedure as in classical analytical chemistry. In order to extract the chemically relevant information, it is necessary to spend a long time on manual data handling. We will develop new data analytical algorithms. These algorithms will allow to automatically extract even very complex overlapping low signal analytes from chromatographic data in a robust, reproducible and timely manner. We claim that current state-of-the-art chromatographic data analysis leaves only a blurry footprint of the information gathered and that modern curve resolution methods have the potential to change that; most notably the tensor method PARAFAC2.

Using methods such as PARAFAC2 requires skilled tensor data analysts. We will develop an expert system, that will allow non-statisticians to critically evaluate complex chemical data based on PARAFAC2. This is where the abductive approach will be crucial for combining the tensor data analysis with chemical insight.

9. Sub-activities in the entire project period

The original parts of the project were to,

1. automatically selecting candidate retention time intervals under the premise that PARAFAC2 works better when using narrow intervals,
2. fitting PARAFAC2 models on individual intervals using distributed and cloud computing to speed up analyses,
3. determining which models are appropriate, and possibly excluding outliers and artefacts extensively using resampling of PARAFAC2 and in various intervals for the same chemical peaks,
4. determining which components represent chemical information and which represent baseline using the chemical knowledge of how well-resolved peaks behave,
5. assessing the best representation of each compound from different candidate intervals to allow selecting the best of many representations of the same peak,
6. attaching a quality assessment to each resolved chemical component to allow the chemist to disregard well-modelled peaks completely as those require no further attention,
7. suggesting a chemical explanation for chemical components that fail step 6,
8. identifying the compounds (using e.g. NIST databases),
9. producing a peak table for further analysis.

All have been achieved (see next section though).

10. Deviations

Scientific:

We have not fully solved the issue of attaching a quality assessment to each resolved chemical (P6) to help the chemist but have an initial prototype assessment for a quality assessment for the elution profile based on a neural network. We included obtained NIST Match Factors (MFs) as a quality control for the obtained mass spectra for each peak. The original plan to “Suggest” a chemical explanation for identification failure (fail step P6) – relies on finding common

chromatographic issues. We have identified several common issues of measurements (overloading, tailing, etc.), but automatically marking these as suspect and accurately handling them is no easy task.

We have partially postponed the idea of a 'Distributed and Cloud' (P2) version of the software PARADISe, instead opting to handle large data locally using out-of-memory techniques. We have also reduced the computational demand by the PARADISe software, making distributed analysis less relevant.

We have not tackled automatic detection and exclusion of outliers and artefact by resampling (P3). This is a computationally expensive process and our efforts have focused on methodical improvements to speeding up PARAFAC2 computation and efficiently run resampling methods.

Financial:

Milk Levy Fund: DKK 963,000

Arla Foods: DKK 798,000

Own funding: DKK 451,000

11. Project results

The mathematical model, PARAFAC2, used to identify chemical profiles, has been substantially improved, i.e. data is now analyzed 10-100 times faster, and it is possible to prevent the spectra from being modelled as non-negative, which is in line with the existing analytical chemistry knowledge.

Selection of intervals (areas where chemical information may be present) is the most time-consuming part for the analyst/user. We have developed a prototype that automates this step; however, still providing feedback to the user and identifying problem areas (where the results are insufficient or uncertain) – allowing the user to remain in control of the process.

We have improved the user interface and the feedback provided by the software based on feedback from users and own insights. The improvements have been implemented in the software PARADISe version 6.0.0, which is available free of charge. More info may be obtained from: https://ucphchemometrics.com/paradise/

Through our collaborations, we have experienced the current state of art first hand and it has become very clear that our new implementation offers huge improvements over the current state of art. Through concrete results within the project, we have really propelled the software implementations forward in terms of analysis speed up, user experience, and reduced user-induced biases. We now have a dedicated set of users that provide valuable feedback, which has allowed us to test our assumptions and improve a roll-out of the software to industry and academia.

12. The relevance of the results, including relevance for the dairy industry

User/dairy-industry oriented:

- Quality of obtained results from chromatographic data analysis
    - More accurate estimation of analyte concentration
    - More analytes observed by deconvolution of overlapping peak elution, and at a lower limited of detection.
- Extreme reduction on time spent on data analysis and/or increase the amount of data analyzed with the same resources.
    - Also reducing time spent by the analysts/users through increased automation.
- Reduced user-induced variability and bias, making the obtained results more reliable.
- Applications

    - Flavour characterization
    - Understanding the maturation process.
    - Storage experiments
- General uses of GC-MS, both targeted and untargeted.

Societal oriented:

The developed method greatly enhances untargeted chemical analysis of cheese and dairy processes. Improvements made in untargeted analysis allow for:

- Better quantification of chemicals affecting dairy-related processes (taste, maturation, oxidation, and storage)
- Discovering new, subtle, or unknown underlying chemical processes affecting the dairy products by lowering the limited of detection.
- More accurate measures of chemical information in the product will facilitate increased understanding of the underlying processes and possibly lead to the development of new products or procedures.

Scientific oriented:

Several important issues have been resolved in the project. One PhD (Huiwen Yu) has worked on improving the numerical algorithms used inside PARADISe and several papers have contributed to the automation and improvement of how to apply the models in practice. In essence, we have provided the skeleton for the first fully automated expert system for creating peak tables. Besides being of huge financial and technological importance, it is also in itself a very interesting scientific achievement.

13. Communication and knowledge sharing about the project

Papers in international journals:

- Baccolo, G., Quintanilla-Casas, B., Vichi, S., Augustijn, D., & Bro, R. (2021). From untargeted chemical profiling to peak tables–A fully automated AI driven approach to untargeted GC-MS. TrAC Trends in Analytical Chemistry, 145, 116451. https://doi.org/10.1016/j.trac.2021.116451
- Yu, H., Augustijn, D., & Bro, R. (2021). Accelerating PARAFAC2 algorithms for non-negative complex tensor decomposition. Chemometrics and Intelligent Laboratory Systems, 214, 104312.
- Yu, H., & Bro, R. (2021). PARAFAC2 and local minima. Chemometrics and Intelligent Laboratory Systems, 219, 104446. https://doi.org/10.1016/j.chemolab.2021.104312
- MA Quelal-Vásconezac, R Macchioni, G Livi; É Pérez-Esteve, MJ Lerma-Garcíad, P Talens, JM Barat, MA Petersen, R Bro, Automatic and non-targeted analysis of the volatile profile of natural and alkalized cocoa powders using SBSE-GC-MS and chemometrics, Food Chemistry, 389, 2022, 133074. https://doi.org/10.1016/j.foodchem.2022.133074
- C Psarras, L Karlsson, R Bro, P Bientinesi, Accelerating Jackknife Resampling for the Canonical Polyadic Decomposition, Frontiers in Applied Mathematics and Statistics, 8, 2022, https://doi.org/10.3389/fams.2022.830270

Easily read papers:

- Bro, R., Augustijn, D., Hinrich, J.L., Yu, H., Petersen, M. A. & Rauh, V. (2021). Få styr på ostens aromaprofil – Kunstig intelligens gør det muligt at automatisere og forbedre kemiske analyser dramatisk. Mælketidende 2021(1). https://maelkeritidende.dk/sites/default/files/udgivelser/Forskningsartikler/start_aroma.pdf

- Hinrich, J. L., Bro, R. & Rauh, V. (2023). Hurtigere og bedre aroma-analyser. Mælkeritidende. Mælkeritidende 2023 (2): p. 10-11. https://maelkeritidende.dk/sites/default/files/udgivelser/Forskningsartikler/sider_fra_maelkeritidende_nr._2_-_2023_-_hoej.pdf

Student theses:

Frederikke Hjertholm Nielsen, Exploring volatile profiles of White wine. Minimizing analytical error and applying exploratory data analysis, June 2022, University of Copenhagen.

Oral presentations at scientific conferences, symposiums etc.:

Multi-way analysis in chemistry, Envisioning Data Driven Advances in Measurements and Instrumentation for Chemical Discovery. Workshop funded by the Chemical Measurements and Instrumentation (CMI) division of National Science Foundation, June 2022, University of Copenhagen.

Huiwen Yu, All-at-once Nesterov-like Extrapolated PARAFAC2-ALS: A Fast and Robust Complex Tensor Decomposition Algorithm for Analyzing GC-MS Data, Annual Conference of the Federation of Analytical Chemistry and Spectroscopy Societies, USA, 12 Oct 2020.

Rasmus Bro, How to make the most of your data, FoodDay, UCPH Food Day – Food & Digitalisation, Food Production Data for All, Aug 26, 2020.

Rasmus Bro, Handling GCMS data, 11th National meeting on chromatography, Caparica, Portugal, Dec 9-11, 2019.

Rasmus Bro, Automating GCMS. Mini-arctic Analysis, 2019, Civita di Bagnioreggia, Italy, Oct 23-24, 2019.

14. Contribution to master and PhD education

We have had several visiting PhD students working on this project. Most notably Huiwen Yu that did his whole PhD on PARADISe and collaborated closely with the project team. Also, Beatriz Quintanilla Casas from University of Barcelona has spent significant time on improving aspects of the PARADISe software.

15. New contacts/projects

- Core Collaborators
    - Michael Agerlin and Pedro Martínez Noguera, University of Copenhagen, Department of Food Science
    - Valentin Rauh, Arla Foods & Peter Skou, Arla Foods Ingredients
    - Beatriz Quintanilla Casas, University of Barcelona, Nutrition and Food Safety
    - Cleo Lisa Davie-Martin and co-workers, University of Copenhagen, Department of Biology, Terrestrial Ecology
- Potential new collaborators
    - Andrea Warburton, University of Otago, Department of Food Science
    - Erica Jaakkola , Lund University, Department of Physical Geography and Ecosystem Science
    - Mircea Martiniuc, University of Glasgow, School of Medicine

There is a huge potential for expanding on the usefulness of PARADISe. It is a significant boost that PARADISe provides for GC-MS. One obvious place to try to also use PARADISe is for LC-MS but there are several developments that are critical to be possible to apply PARADISe on such data.