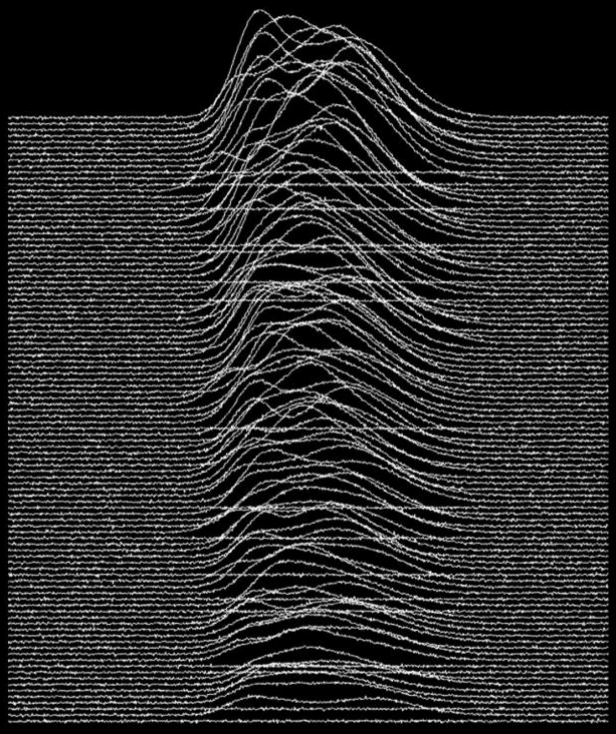
UNIVERSITY OF COPENHAGEN

FACULTY OF SCIENCE
DEPARTMENT OF FOOD SCIENCE

PHD THESIS

PAUL-ALBERT SCHNEIDE



New algorithms and perspectives for information extraction from gas and liquid chromatography coupled to mass spectrometry data

SUPERVISOR: RASMUS BRO

New algorithms and perspectives for information extraction from gas and liquid chromatography coupled to mass spectrometry data

PhD thesis

Paul-Albert Schneide

This thesis has been submitted to the PhD School of the Faculty of Science

University of Copenhagen

Department of Food Science

Rolighedsvej 26, 1958 Frederiksberg C

Denmark

Title	New algorithms and perspectives for information extraction from gas and liquid chromatography coupled to mass spectrometry data.
Author	Paul-Albert Schneide
Submission Date	1st of September 2024
Deference	12th of November 2024
Principal supervisor	Rasmus Bro, Professor, Department of Food Science, Faculty of Science, University of Copenhagen
Co-supervisor	Miriam Mathea, Global Scientific Discipline Lead Cheminformatics, BASF SE

Assessment Committee

Chairman	Frans W.J. van den Berg, Associate Professor, Department of Food Science, Faculty of Science, University of Copenhagen
External	Jose M. Amigo, Professor, Department of Analytical Chemistry, University of Basque Country
External	Robert E. Synovec, Professor, Department of Chemistry, University of Washington

Acknowledgements

This PhD work has been funded by an internal research project at the Department of Analytical Science within BASF SE. I would like to thank BASF SE and more specifically the people that created this great opportunity for me: Kathrin Wolter, Grit Baier, Miriam Mathea, Christina Köppen, and Anke Reinold, among many others. I would also like to thank my colleagues and former colleagues at BASF SE, and especially, Emma, Rafael, Stefan, and Rebeca for their great support in balancing my responsibilities at BASF SE with my PhD work. It was Anke, who encouraged me to do the PhD, and Miriam who agreed to supervise me during this time. Thanks to both of you for great mentorship during the past years of my career!

Doing a PhD has been humbling in the beginning and stressful toward the end, but nonetheless, it was a great adventure. I want to thank the many people that made this one of the most exciting periods of my life. I would like to thank all my co-authors, Rasmus, Neal, Michael, and Oskar for many inspiring conversations and the great ideas you have enriched my research with. I would also like to acknowledge the important discussions with Frans and Giorgio that improved my understanding of the nuances in signal processing, and I thank both of them for taking the time to teach me. I am grateful to Prof. Frank Glorius and the team from ChemInnovation for hosting me and making me feel welcome during my two-month exchange at the University of Münster. Certainly, without the enthusiasm, creativity, openness, and positivity of Rasmus this PhD would not have been the same. I have learned so much from you, Rasmus, academically and personally. You have been the best supervisor I could have hoped for.

This PhD has also enriched my life because I have met many colleagues who became close friends. I would like to thank Ching for many good conversations about academia and life, and for making me laugh at myself. I am grateful for having met Pedro, Pablo, Tjark, Maxime, and Bogomil, thank you for so many joyful and unforgettable times! I would like to high-five my fellow chemometricians, friends, and conference buddies Beatriz and Oksana for many great moments during our travels.

Finally, I want to thank my family, friends, and my girlfriend for having my back and supporting me with their love and affection. Thank you, Cecilie, for being patient with me when I am not patient with myself, for making me laugh, and for your empathy and love during the past years, and especially during the end of my PhD.

Abstract (English)

Chromatographic separation with mass spectrometric detection is one of the most powerful instrumental combination analytical chemistry has to offer. The data acquired from these instruments are information-rich but complex. Therefore, it is still a challenge to convert raw data into actionable, chemical information. Consequently, data analysis often becomes the most time-consuming step in the whole analytical process. This is even more the case for untargeted measurements, in which the goal is to obtain a comprehensive characterization of complex samples. These types of measurements are increasingly popular in environmental science, metabolomics, food science, and the chemical industry.

A lot of research effort has been devoted to developing data analysis capabilities in the form of open-source or commercial data analysis workflows, and software packages. While these efforts have facilitated the conversion from raw data into chemical information, a satisfactory point has thus far not been reached. Existing data analysis workflows require users to define many parameters that have a strong influence on the data analysis, limit the reproducibility of analytical results and may lead to incorrect conclusions. More abstractly, the reason for this overparameterization in existing workflows can be attributed to the reliance on heuristics and a lack of structural assumptions. Furthermore, most open-source data analysis workflows can only process one-dimensional chromatographic data. However, in the field of chemometrics, different methods based on multivariate curve resolution or tensor decomposition have been developed that incorporate *a priori* knowledge about the chromatographic data. Applications of these methods to one- and two-dimensional chromatography have been demonstrated. Nevertheless, there are also some drawbacks to these chemometric methods. Specifically, multivariate curve resolution methods suffer from rotational ambiguity, which means that they do not provide a unique solution. Multilinear tensor decomposition methods, on the other hand, provide unique solutions but often have assumptions that are too rigid to match the complexity of the chromatographic data. Additionally, applications of curve resolution or tensor decomposition for the analysis of trace-level compounds can be challenged by low signal-to-noise ratios.

In this PhD thesis, two novel approaches are introduced to address the limitations of existing chemometric methods for chromatographic data analysis. The first set of contributions introduces different versions of shift-invariant models incorporating trilinearity (Paper 1), soft-trilinearity (Paper 2), and multi-linearity (Paper 3) to overcome problems with rotational ambiguity. Additionally, they provide a more flexible approach compared to traditional

multilinear tensor decomposition methods, better accommodating the complexity of chromatographic data. Applications of these models are demonstrated to one- and two-dimensional gas chromatography coupled with mass spectrometry. The second contribution, detailed in Paper 4, presents a signal processing workflow specifically designed for trace-level suspect screening in two-dimensional chromatography coupled with high-resolution mass spectrometry. This workflow includes a mass filtering algorithm that efficiently extracts pure mass spectra, even in complex samples with low signal-to-noise ratios.

Abstract (Danish)

Kromatografisk adskillelse kombineret med massespektrometri er blandt de stærkeste instrumentelle værktøjer i den analytiske kemi. Data udvundet med disse metoder er informationsrig men ligeledes kompleks. Det er derfor stadig er en udfordring, at omsætte rå data til anvendelig kemisk information, i en grad hvor dataanalysen kan være det mest tidskrævende trin i hele analysen. Dette gælder især for forsøg hvor målet er en omfattende karakterisering af sammensætningen af komplekse prøver. Denne type målinger bliver i stigende grad populær inden for fag som miljøvidenskab, metabolomics, fødevarevidenskab samt i den kemiske industri.

En betydelig forskningsindsats danner grundlag for udviklingen af dataanalysemuligheder i form af open source- eller kommercielle dataanalyseteknikker og softwarepakker. Selvom dette arbejde har muliggjort en omdannelse af kromatografiske rå data til anvendelig kemisk information, er resultatet endnu ikke fuldt tilfredsstillende. Eksisterende dataanalyseteknikker kræver, at brugeren definerer mange parametre, hvilket har en betydelig indvirkning på analysekvaliteten, herunder en nedsat reproducérbarhed og endda falske slutninger. Det er især den heuristik-baserede tilgang og manglen af datastrukturelle antagelser, som ligger til grund for den omtalte overparameterisering. En yderligere begrænsning ligger i, at de fleste open source analyseteknikker kun kan anvendes på endimensionelle kromatografiske data.

Inden for kemometrien er der udviklet metoder baseret på Multivariate Curve Resolution eller ”tensor decomposition” som netop tager udgangspunkt i *a priori* viden om kromatografiske data. Brugen af disse metoder til en- eller todimensionel kromatografi er tidligere demonstreret. Dog er der også ulemper ved metoderne. Disse ulemper omfatter blandt andet rotationsambiguitet i Multivariate Curve Resolution, hvor metoden ikke kan give ét unikt resultat. Multilineær tensor decompostion giver derimod ét unikt resultat, men kommer med for rigide antagelser set i forhold til kompleksiteten bag de kromatografiske data. Endvidere, ved analyse af indholdsstoffer på sporingsniveau kan brugen af curve resolution eller tensor decomposition udfordres af et lavt signal-støj-forhold.

I denne Ph.D.-afhandling introduceres to nye tilgange til at omgå begrænsninger i de eksisterende kemometriske metoder for kromatografisk dataanalyse. Det første bidrag fremlægger forskellige versioner af ”shift-invariant” modeller som inkorporerer ”trilinearity” (Artikel 1), ”soft-trilinearity” (Artikel 2) og ”multi-linearity” (Artikel 3) som metoder til at overkomme problemer med rotationsambiguitet. Derudover er modellerne mere fleksible, sammenlignet med traditionelle multilineære tensor decomposition metoder, og derfor bedre til

at omfavne kompleksiteten af de kromatografiske data. Brugen af disse modeller bliver vist for en- og todimensionelle data fra gaskromatografi-massespektrometri. Det andet bidrag, beskrevet i Artikel 4, omfatter en signal-processeringsteknik udviklet til at screene for definerede komponenter på sporingsniveau med todimensionel kromatografi kombineret med højtopløst massespektrometri. Teknikken inkluderer en massefiltrerings algoritme, som effektivt udvinder rene massespektre, selv i komplekse prøver med lavt signal-støj-forhold.

List of scientific contributions

Paper 1

Schneide P-A, Bro R, Gallagher NB. Shift-invariant tri-linearity—A new model for resolving untargeted gas chromatography coupled mass spectrometry data. Journal of Chemometrics. 2023; 37(8):e3501. doi:10.1002/cem.3501

Paper 2

Schneide P-A, Gallagher NB, Bro R. Shift invariant soft trilinearity: Modelling shifts and shape changes in gas-chromatography coupled mass spectrometry. Chemometrics and Intelligent Laboratory Systems. 2024; 251. doi: 10.1016/j.chemolab.2024.105155.

Paper 3

Schneide P-A, Armstrong MS, Gallagher NB, Bro R. Unlocking new capabilities in the analysis of GC×GC-TOFMS data with shift-invariant multi-linearity. Journal of Chemometrics (submitted)

Paper 4

Schneide P-A, Kronik OM. A signal processing workflow for suspect screening in LC×LC-HRMS: Efficient extraction of pure mass spectra for identification of suspects in complex samples using a mass filtering algorithm. Analytical Chemistry (submitted)

List of Abbreviations

^{1}D	First separation dimension of a GC×GC or LC×LC measurement
^{2}D	Second separation dimension of a GC×GC or LC×LC measurement
AF	Amplitude Filter
AI	Artificial Intelligence
ALS	Alternating Least Squares
AMDIS	Automated Mass Spectral Deconvolution and Identification System
APCI	Atmospheric Pressure Chemical Ionization
CAMERA	Collection of Algorithms for Metabolite Profile Annotation
CCE-MSP	Class Comparison Enabled Mass Spectrum Purification
CI	Chemical Ionization
COW	Correlation Optimized Warping
DAD	Diode Array Detection
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
DTW	Dynamic Time Warping
EFA	Evolving Factor Analysis
EI	Electron Ionization
EIC	Extracted Ion Chromatogram
ESI	Electron Spray Ionization
FDA	Food and Drug Administration
FID	Flame Ionization Detector
FI	Field Ionization
FF	First Derivative Filter
GC	Gas Chromatography
GC×GC	Comprehensive two-dimensional Gas Chromatography
GNPS	Global Natural Product Social Molecular Networking
GRAM	Generalized Rank Annihilation Method
HPLC	High Pressure Liquid Chromatography
ICH	International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use
IM	Ion Mobility spectrometry
LC	Liquid Chromatography

LC×LC	Comprehensive two-dimensional Liquid Chromatography
LLE	Liquid-Liquid Extraction
MASST	Mass Spectrometry Search Tool
MCR	Multivariate Curve Resolution
m/z	Mass-To-Charge Ratio
MIR	Mid Infrared radiation
MS	Mass Spectrometry
MS/MS	Tandem mass spectrometry
MS^1	Precursor ion measurement in the first stage of a tandem mass spectrometry measurement
MS^2	Product ion measurement in the second stage of a tandem mass spectrometry measurement
MS^E	Alternative term for MS^2 if product ions spectra have been measured for all precursor ions (DIA all ion fragmentation)
MSI	Metabolomics Standards Initiative
NIR	Near Infrared radiation
NMF	Non-negative Matrix Factorization
NMR	Nuclear Magnetic Resonance spectroscopy
PARADISe	PARAFAC2-based Deconvolution and Identification System
PARAFAC	PARAllel FACtor analysis
PARAFAC2	PARAllel FACtor analysis 2
PARASIAS	PARAFAC Applied to Shift Invariant Amplitude Spectra
PCA	Principal Component Analysis
PLS-R	Partial Least Squares Regression
QA	Quality Assurance
QC	Quality Control
QTOF	Quadrupole-time-of-flight tandem mass spectrometer
RANSAC	Random Sample Consensus algorithm
ROI	Region Of Interest
SF	Second Derivative Filter
SIM	Selected Ion Mode
SIML	Shift Invariant Multilinearity
SIST	Shift Invariant Soft-Trilinearity
SIT	Shift Invariant Trilinearity

SNR	Signal-to-Noise Ratio
SPE	Solid-Liquid Extraction
SVD	Singular Value Decomposition
TIC	Total Ion Chromatogram
TOF	Time-of-flight mass analyzer
UV	Ultraviolet radiation

Notation

$\underline{\mathbf{X}}^{(p)}$	*pth* order tensor
$\mathbf{X}$	$(I \times J)$ matrix, equivalent to $\underline{\mathbf{X}}^{(2)}$
$\mathbf{x}$	$(I \times 1)$ vector, equivalent to $\underline{\mathbf{X}}^{(1)}$
x	scalar, equivalent to $\underline{\mathbf{X}}^{(0)}$
$\mathbf{x}^{\mathrm{T}}$	$(1 \times I)$ transpose of $\mathbf{x}$
$\|\| \cdot \|\|_F^2$	Frobenius norm
$\hat{f}(k)$	Fourier transform of $f(n)$
$\hat{f}^*(k)$	Complex conjugate of $\hat{f}(k)$
$\|\hat{f}(k)\|$	Amplitude spectrum of $f(n)$

Table of contents

1 Introduction

Different definitions exist of what modern analytical chemistry as a field is, and what roles and responsibilities analytical chemists have.[1–4] However, all definitions agree that analytical chemistry plays a crucial role in science, industry, and society, because it provides information that help scientists getting a better understanding of the material world and further help people and systems to make well-informed decisions.

In 1981 Bruce Kowalski stated the importance of "analytical chemistry as an information science" and foresaw that the role of analytical chemists in the future would be the one of "solution providers" rather than mere "data generators".[5] Further, Kowalski concluded that this transition within analytical chemistry required data analysis tools which can efficiently extract chemical information out of datasets of increasing size and complexity.[5] The development and application of data analysis methods and algorithms that extract chemical information from raw data is at the heart of chemometrics.[6] Many of the perspectives that Bruce Kowalski has put forward are today, more than 40 years later, still valid. It is even more evident today, how much there is to gain, if feedback loops based on chemical information would be faster. This is why large initiatives in the chemical, pharmaceutical, and food industries have kicked off for online process monitoring using process analytical technology. These initiatives aim to improve process stability, shorten development times, and increase process safety.[7–9] Fast spectroscopic sensors like UV, NIR, MIR or Raman are used in combination with chemometric models to replace time consuming offline analysis.[7–9]

Outside production the hunger for faster access to comprehensive chemical information is huge, as well. For instance, for the identification of biomarkers in metabolomics[10,11], understanding polymer waste quality for chemical recycling[12], monitoring substances of concern in the environment[13,14] or for authenticity testing and development of food products.[15] The main source of chemical information in these use cases are chromatographic instruments with multi-channel detectors. Especially, chromatography coupled to mass spectrometric detectors display maybe the most important family of analytical instruments in this context. One way to find answers to these increasingly difficult analytical questions and to address the desire for faster access to comprehensive chemical information is the development of improved hardware. Fast chromatography methods in combination with time-of-flight mass spectrometers can speed up measurement times significantly, while providing comprehensive chemical information – in principle.[16,17] Furthermore, comprehensive two-dimensional chromatography methods can separate highly complex chemical mixtures in a single measurement.[18–22] However, acquiring

data faster and in more complex data structures ultimately shifts the bottleneck from the data generation to the data analysis step. In this regard, the development of chemometric methods, such as curve resolution and tensor decomposition, addressed some fundamental problems related to the analysis of complex chromatographic datasets.[23–26] Nevertheless, some challenges with respect to algorithms, model assumptions and the complexity of chromatographic data remained.[27–31] Outside the chemometrics community, comprehensive data analysis tools such as XCMS, MZmine 3, and MS-DIAL have been developed for one-dimensional chromatography hyphenated to mass spectrometry data.[32–35] These tools offer end-to-end workflows for data analysis and have gained widespread popularity.[36–38] Recently, studies have criticized the lack of transparency and reproducibility concerning the results obtained using these tools.[39–43] For two-dimensional chromatography, data analysis tools are less abundant, and lacking data analysis capabilities are considered a hindrance to the broader adoption of the technology.[44–46]

Hence, data analysis, and especially the extraction of chemical information from complex chromatographic raw data, remains an ongoing challenge to which this work attempts to propose some solutions.

1.1 Aim and outline of the thesis

On a more general note, the purpose of the thesis is to make a small contribution to Bruce Kowalski's vision of analytical chemistry as an information science. Specifically, the aim is to investigate existing approaches and to develop complementary methods and algorithms for the extraction of chemically meaningful information from chromatography hyphenated to mass spectrometry data.

The investigation part starts with an overview over different analytical strategies and why to use any of them (*Chapter 2: Targeted, untargeted, and semi-targeted analysis strategies*). Next, chromatographic methods coupled with mass spectrometry for detection are introduced. The role of artifacts is especially emphasized, as they play a crucial role in constructing a hypothesis about the chromatographic data structure (*Chapter 3: Chromatography, mass spectrometry, and raw data pre-processing*). After establishing a theoretical foundation for understanding the structure of chromatographic data, the use of curve-resolution and tensor decomposition methods for the extraction of chemical information is discussed (*Chapter 4: Information extraction using curve-resolution and tensor decomposition methods*). Alternative data analysis methods and workflows are presented, and it is discussed to which extent these can be combined with or complemented by curve resolution and tensor decomposition methods (*Chapter 5: Alternative methods for information extraction*). In the final part, main results of the investigation are summarized, and future development directions and perspectives are discussed (*Chapter 6: Conclusion and future perspectives*).

The development of new algorithms and methods is captured in four scientific articles appended to the thesis. In summary, the newly proposed methods extend existing methods by being computationally more efficient, allowing a close approximation to experimentally observed data structures, and improving resolution capabilities in the sense that accurate information can be recovered from low signal-to-noise data. In [Paper I], a novel and highly efficient algorithm for shift-invariant trilinearity (SIT) is introduced and the application to GC-MS data is demonstrated. In [Paper II], an extension of the SIT algorithm called shift-invariant soft trilinearity (SIST) is proposed, which can more flexibly handle chromatographic artifacts like peak shape changes. In [Paper III], an extension of the SIT algorithm which can model multilinear data structures with shifts in more than one mode is proposed for modeling comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC×GC-TOFMS) data. In [Paper IV], a new algorithm for the extraction of

clean mass spectra from comprehensive two-dimensional liquid chromatography data coupled to high resolution tandem mass spectrometry (LC×LC-MS/MS) data is described for suspect screening in wastewater samples. Furthermore, limitations of multivariate curve resolution alternating least squares (MCR-ALS) for the extraction of mass spectra of trace-level compounds in LC×LC-MS/MS are discussed.

Many aspects crucial to moving analytical chemistry toward an information science are not covered or only superficially discussed in this work. For starters, information extraction is only one part of the analytical data analysis pipeline, as it additionally comprehends raw data pre-processing, diagnostic checks, and downstream data analysis. Moreover, a developed algorithm is not yet a tool that can readily be used but rather needs to be implemented into a software platform or data analysis pipeline to be useful for analytical chemists and practitioners. Although methods described in [Paper I-II] have been implemented in software platforms[47,48], this thesis does not provide any details about the difficulties of building functional and user-friendly software or data analysis pipelines. Further, all aspects concerning laboratory automation, information management systems, data standards and data models (not to be confused with data structures), are outside the scope of this work. However, it has been recognized that many important initiatives have been started in science and industry which hopefully will leverage the use and facilitate the implementation of the developed methods for practical application.[49–51]

2 Targeted, untargeted, and semi-targeted analysis strategies

Good chemometricians try to solve practical problems. Hence, it is crucial to understand the goal and ambition behind an analytical measurement. It is crucial that both, the analytical chemist and the chemometrician have the same understanding regarding the goal to be achieved, in order to select the appropriate chemometric tools, methods and, algorithms. For example, to investigate the difference in the chemical fingerprints of two sets of samples, describes a completely different task compared to investigating which chemicals between two sets of samples differ. While the first task can be achieved with an unsupervised method and moderate effort, the latter one requires comprehensive mining of (semi-)quantitative chemical information followed by rigorous statistical testing.

This Chapter will present and discuss different experimental set ups that are used in chromatography. The goal is to describe the differences between a targeted, semi-targeted and untargeted analysis approach. In **Figure 1** essential differences between the methods are summarized. Furthermore, arguments for using any of the listed approaches as an analytical strategy are discussed. The distinction between these three analyses strategies, developed alongside technological and methodological improvements in mass spectrometry, became especially popular within metabolomics and other emerging omics-fields.[52–55] Chapter 5 and 6 will mainly focus on methods for extracting information from data acquired in semi-targeted and untargeted experiments.

2.1 Targeted analysis

Targeted analysis is the "classical" analytical procedure of quantifying a selected number of known analytes in a specified sample matrix. This approach is widely employed, e.g., in chemical, pharmaceutical and food industry to test if specifications of manufactured goods with respect to the allowed levels of impurities are fulfilled. Other applications can for instance be found in environmental monitoring, e.g., quantification of specific chemical pollutants in complex samples.[56,57] Comprehensive guidelines published by regulatory authorities and councils like the *Food and Drug Administration* (FDA) and the *International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use* (ICH) provide a framework for the development and validation of analytical procedures for targeted analysis.[58,59] For the development of a targeted method, it is important to first state the intended purpose of its use and then demonstrate that the developed method is fit for purpose in a

validation step.[60–62] When describing the purpose of an analytical procedure, the analytes to be analyzed must be defined together with the expected concentration range of the analytes, the sample matrix, and the required selectivity, and accuracy.[58] To reduce the effect of disturbing matrix constituents and to pre-concentrate the analytes of interest, selective sample preparation techniques like liquid-liquid extraction (LLE) and solid-phase extraction (SPE) are used as part of the analytical procedure.[63] Analytes are identified and quantified using e.g., isotopic labelled reference standards.[54,56] Additionally, quality assurance procedures need to be implemented to monitor and correct for variations in the analytical procedure that are caused by a lack of control over the experimental conditions (e.g., instrument drift).[58,59] However, quality assurance procedures are not only a standard requirement for targeted analysis but should also be part of semi-targeted and untargeted analysis protocols.[64,65]

Targeted analysis is an excellent strategy for testing analytical hypotheses such as: *a pharmaceutical product fulfills specifications*[66], *treatment increases the level of a specific metabolite*[63], *human industrial activity leads to increased levels of specific organic pollutants*[56,57]. However, the utility of targeted analysis is limited when it comes to the generation of new hypotheses.[67] This is, because the analytical scope of a targeted analysis, defined by the number of analytes monitored, is limited. In fact, maximizing the selectivity of the analytical procedure toward the targeted analytes likely means being "analytically blind" for other potentially relevant compounds.[54]

2.2 Untargeted analysis

Untargeted analysis (or non-targeted screening) is a method used for the broad qualitative characterization of large sample sets by comparing relative concentrations of annotated compounds between samples.[68–70] Conversely to targeted analysis, the identity of compounds is usually not verified using reference standards due to economical and practical restrictions.[54,71] Instead, extracted mass spectral information and retention times of compounds are compared against databases for annotation.[68] The peak areas of compounds can be used as relative concentrations, after normalizing the data for matrix and batch effects (e.g., using compound class specific internal standards).[54,71] To avoid discrimination of analytes and introducing bias, sample preparation is often reduced to a minimum compared to targeted analysis.[54] This may lead to analytical challenges, which will be discussed in detail in Chapter 3.

Depending on the study design, untargeted analysis can further be divided into supervised and unsupervised untargeted analysis. Supervised untargeted analysis refers to the situation in which a class structure (patients vs. control, contaminated vs. reference) is *a priori* known.[72,73]

Early applications of untargeted analysis using GC-MS and LC-MS for environmental monitoring have already been reported in the late 1970s.[74,75] Since then, untargeted analysis found a lot of attention in environmental sciences[76,77], the metabolomics community[78,79], food science[80,81], and beyond. It needs to be highlighted that differences in the analytical scope of untargeted analysis exist between communities. For instance, while it might be acceptable (or more accepted) for metabolite discovery to focus on higher abundant analytes[67,70], organic pollutants relevant for environmental monitoring are often present at a trace level.[82]

Hollender et al. describe that, based on estimates by Schwarzenbach et al., between 30,000 to 70,000 chemicals (including pharmaceuticals and surfactants) are used in households alone, of which only a very small subgroup (< 100) is included in targeted regulatory monitoring programs.[69,83] Therefore, targeted analysis alone is insufficient for providing a comprehensive overview of the fate of these emitted chemicals and their impact on the environment, underlining the importance of untargeted analysis approaches. Untargeted analysis is especially valuable for the discovery of unexpected relationships and the generation of new hypotheses.[70]

The analytical data acquired in untargeted analysis is very complex to the extent where specialized algorithms and data analysis workflows have been developed to deal with this burden. [35,37,84–89] However, data analysis workflows and capabilities are less mature for two-dimensional chromatography, although these techniques are powerful when it comes to the separation of complex biological or environmental samples.[18,44,46,90]

An important finding, that several inter-laboratory studies have shown, is that reproducibility of untargeted studies still needs to be improved to decrease the risk of misleading conclusions.[91–95] Further, it has been shown that GC-MS and LC-MS provide complementary results and hence should be used together in untargeted studies to achieve a high analytical coverage.[91,93] One important lever for improving reproducibility are standardized protocols and guidelines which have been proposed by the *Metabolomics Standards Initiative* (MSI) and more recently by the NORMAN network.[71,96] It needs to be emphasized that standardization should not only cover experimental and analytical procedures but also the data analysis workflows.[97,98]

Although, the development of easy-to-use software and comprehensive data analysis workflows has greatly facilitated the analysis of complex untargeted data (for one-dimensional chromatography), it has been demonstrated that using different software tools may lead to different analytical results.[39–42] A more detailed discussion on algorithms implemented in different data analysis workflows is the topic of Chapter 5.

2.3 Semi-targeted analysis

Semi-targeted analysis, also referred to as *suspect screening*, conceptually builds the bridge between targeted and untargeted analysis.[99] It is the analysis strategy to employ if the analytical scope can be constrained to a limited chemical space. One example may be that a list of several hundred or even thousands of organic pollutants exist which should be monitored in water samples.[100] The information in a suspect list can comprehend the molecular mass, retention indices, characteristic mass fragments or even reference mass spectra.[71] This *a priori* information might be useful to optimize the analytical procedure toward the relevant chemical space spanned by the suspect compounds.[101,102] While it is often still practically not feasible to obtain reference standards for all suspected compounds, it might be possible to find a group of analytical standards that are representative for the molecule classes on the suspect list. This can improve the reliability of semi-quantitative results in semi-targeted analysis compared to untargeted analysis.[103] More often the aim in suspect screening is the identification of trace-level compounds in complex sample matrices.[100,102,104] Thus, for the identification of the suspected compounds it is required to extract clean mass spectra of low abundant compounds, which can be challenging with state-of-the art data analysis workflows.[82] In [Paper IV], a workflow has been proposed for leveraging suspect compound information for mass filtering combined with MCR-ALS to extract clean mass spectra from trace-level compounds. The proposed workflow significantly improved compound identification in LC×LC-MS/MS measurements of a wastewater sample using mass spectral databases.[105]

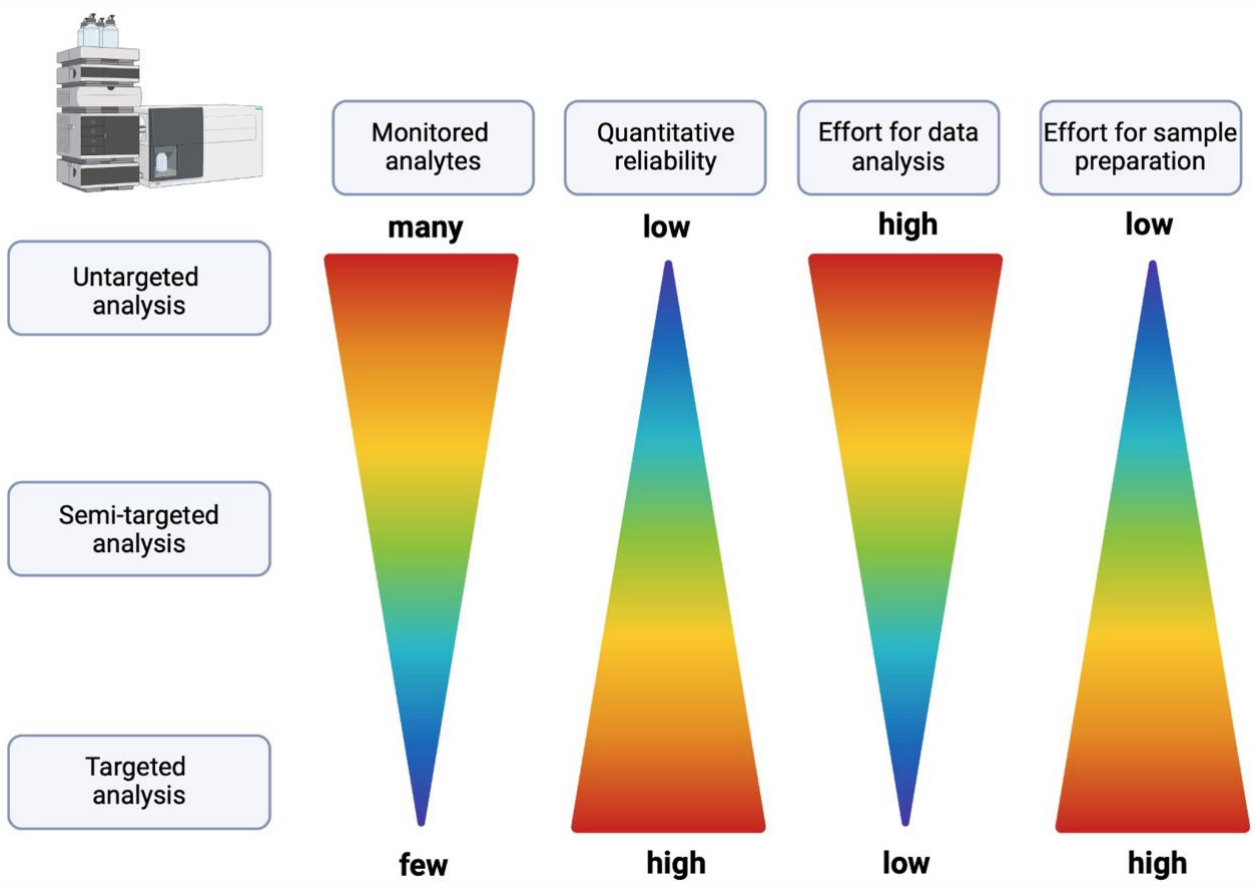


***Figure 1**: Contrasting the differences between targeted, semi-targeted and untargeted analysis approaches. Created with BioRender.com*

3 Chromatography, mass spectrometry, and raw data pre-processing

The use of a chemometric model should be justified by some a priori hypothesis regarding the structure of the modelled data. For example, the use of partial least squares regression (PLS-R)[106] for modelling quantitative relationships in NIR data is often justified by the applicability of the Beer-Lambert Law[107], and the known electro-magnetic characteristics of NIR spectra. Thus, the hypothesized data structure justifies the use of a certain chemometric model.

This Chapter provides the theoretical foundation necessary for discussing chemometric models and chromatographic data structures (GC-MS, LC-MS/MS, GC×GC-TOFMS, LC×LC-MS/MS) in Chapter 4. While some of the presented theory may be familiar to chromatography experts, it is not necessarily common knowledge for chemometricians. Additionally, the Chapter examines differences among mass spectrometric detectors coupled to GC and LC, exploring their potential influence on chromatographic data structures. Although not as exhaustive as specialized textbooks, this theoretical overview aims to explore which artifacts can influence chromatographic data structures, a topic of significant interest in the chemometric community.

The domain-specific terms used in this Chapter are based on the recommended nomenclature for chromatography by the IUPAC[108], on the IUPAC Compendium of Chemical Terminology[109], or are defined in the text. Some clarification regarding the usage of the terms “analyte”, “compound”, and “solute” may be required. The terms “analyte” or “compound” refer to the chemically pure constituents of a sample, while “solute” describes an analyte or compound undergoing chromatographic separation.

3.1 History of chromatography and chromatographic instruments

Chromatography fundamentally describes the process of separating a mixture of compounds, solved in a mobile phase, by their interaction with an immobile, or stationary phase.[110] In this work the term chromatography will exclusively describe applications of column chromatography, in which the mobile phase is pumped through a column containing the stationary phase.[110] Other types of chromatography like planar chromatography are less abundant today but may still be relevant for some applications.[111]

The discovery of the general principle of chromatography dates to the early 1900s and the work of the botanist M.S. Tswett who used liquid chromatography to separate and study plant

pigments.[112] Hence, this first application was also defining the name "chromatography", which literally translates to "color writing" coming from the combination of the Greek words *chroma* and *graphein*. The first chromatography instruments were developed 50 years later by A.J.P. Martin and A.T. James[113,114] enabling the use of gas chromatography as an analytical technique. This technological breakthrough was based on the work from Martin and R.L.M. Synge on partition chromatography from the early 1940s.[115,116] The development of instrumental analytical liquid chromatography systems took a longer time due to the difficult technical realization. The theoretical foundations for such an instrument had been described already by the work of Martin and Synge, though.[115,116] Though it was C. Horvath and S. Lipsky who are recognized as the scientists behind the first prototype of what is now considered one of the most powerful and widely used analytical techniques today – high performance liquid chromatography (HPLC).[117,118] The limitation of gas chromatography being only applicable to the analysis of volatile compounds led to the development of HPLC. Since the only form of liquid chromatography discussed in this work is high pressure liquid chromatography, the abbreviation "LC" will be used equivalently to "HPLC". The application range of LC comprehends a large chemical space including large and polar molecules, which are only accessible to GC after complex sample preparation like derivatization.[119] Nevertheless, GC is still of enormous importance, complementing LC in many applications.[91,93,95] However, the growing demand for handling highly complex samples has driven the development of chromatographic systems capable of separating a larger number of analytes. This ultimately led to the development of the first functional, comprehensive two-dimensional gas and liquid chromatography (abbreviated with GC×GC and LC×LC) instruments in the early 1990s.[120–123] Since then, two-dimensional chromatography as a technology has more matured and found many applications, also outside academia.[18,21,124,125] **Figure 2** provides a schematic overview of the essential parts of GC, GC×GC, LC, and LC×LC systems.

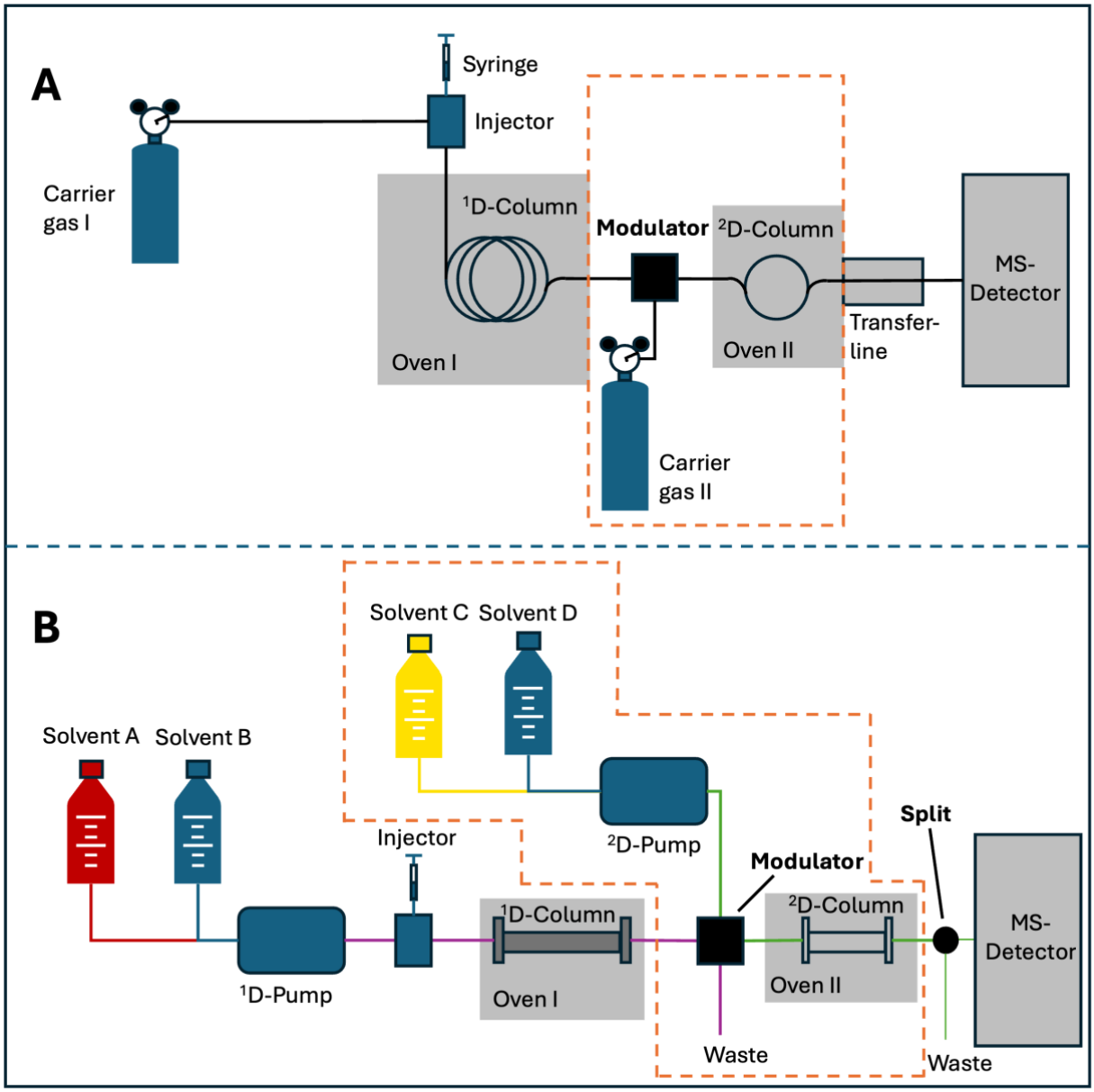


***Figure 2**: Schematic visualization of gas chromatography (**A**) and liquid chromatography (**B**) with mass spectral detection. The parts inside the orange-dashed line highlight the differences between one-dimensional and two-dimensional chromatography. Samples are injected in the injector and the solutes travel through the chromatography column(s) until they reach the detector. The modulator is the heart of comprehensive, two-dimensional chromatography, linking both columns and ensuring continuous (online) injection of fractions from the first column into the second column. The polarity of the ^{1}D- and ^{2}D-column are usually different. This is also the case for the polarity of the mobile phase in two-dimensional liquid chromatography. The separation on the ^{2}D-column in GC×GC and LC×LC is faster than on the first column, in order to avoid under-sampling of the signal in ^{1}D. The mass spectral detection will be discussed in more detail in Section 3.3.*

3.2 Retention mechanism, performance, and artifacts

Throughout the chromatographic process, the solutes are distributed between two phases, a stationary phase, and a mobile phase. In this work the focus will be on chromatographic systems that use either a gas or a liquid as mobile phase. Applications of supercritical chromatography gain in importance but are beyond the scope.[126,127] Further, the stationary phases considered are wall coated capillary columns for GC and reversed phase LC.[128,129]

The extent to which the "traveling time" of a solute through the column (retention time) is delayed by its interaction with the stationary phase depends on the physical and chemical properties of the solute, the stationary phase, and the mobile phase. Additionally, factors such as temperature, pressure, flow rate, and pH also influence the retention time. Some of these influencing factors are used in modern chromatography as parameters to optimize and control the separation process, i.e. by running a temperature program in GC[130,131] and by gradually changing the composition of the mobile phase in LC.[132,133] However, all these factors are potential sources of variability to the signal. The partition process between the stationary and mobile phase is visualized in **Figure 3**. The solutes exiting the column are recorded by a detector, which ideally produces a signal with an intensity proportional to the amount of solute arriving within a given time frame. Detector saturation or matrix effects, discussed in Sections 3.4.2 and 3.4.3, can distort this proportionality. Furthermore, multi-channel detectors such as UV or MS provide a solute-specific response that can aid in identifying the chemical structure. Since untargeted analysis almost exclusively uses MS detectors, these will be discussed in more detail in Section 3.3.

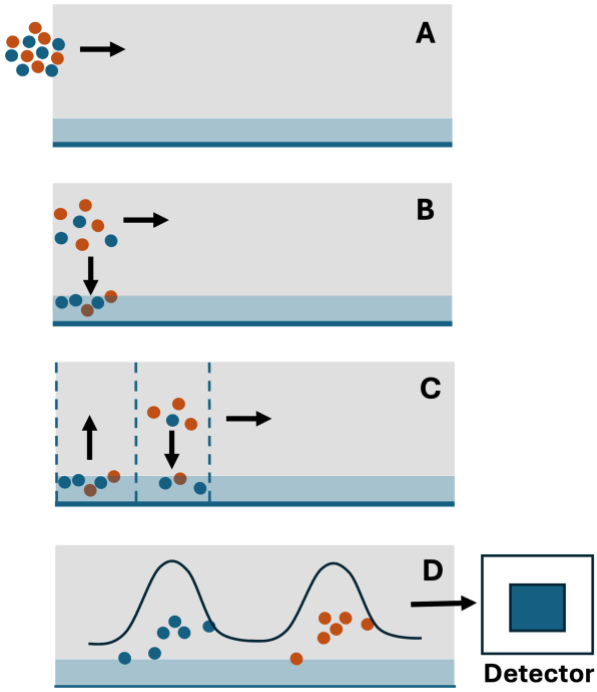


Figure 3: *Separation along a chromatographic column.* ***A****: Mixture of two solutes is injected into the chromatographic system and is at the beginning entirely dissolved in the mobile phase (grey background).* ***B****: Based on their physical and chemical properties, the individual solutes (blue and orange dots) start to interact with the stationary phase (light-blue background).* ***C****: Along the column, the solutes interact multiple times with the stationary phase. This process is described by the concept of theoretical plates (denoted by vertical dashed lines).* ***D****: By interacting with the stationary phase, the solutes experience different retardation and are separated along the column. The solutes leaving the column reach a detector which records a signal that is proportional to the amount of the respective solute arriving in a given time frame and can be specific to the type of solute.*

The result of a chromatographic measurement is a chromatogram, which shows the detector response over the time of the chromatographic measurement. **Figure 4A** shows a simplified version of a gas chromatogram with three distinct peaks representing three different compounds that have reached the detector at different time points. **Figures 4B-E** show overlays of the chromatogram depicted in **Figure 4A** with other measurements (orange colored). The orange-colored chromatograms differ in all cases from the blue-colored chromatograms, which is due to chromatographic artifacts. Specifically, the artifacts can be characterized as retention time shift (**Figure 4B**), band or peak broadening (**Figure 4C**) and peak skewness (**Figure 4D**). Artifacts will be referred to as changes in the chromatograms that occur relative to a reference. For instance, the method that was used to acquire the data shown in **Figure 4C** can generate chromatograms with tight peaks (blue line) but the widths of the peaks changed across a set of measurements (orange line). In **Figure 4E**, a fourth peak appears in close proximity to the third peak resulting in overlapping, or co-eluting signals. Although the problems of band broadening and co-elution are somewhat related problems (smaller bands will decrease the probability of co-elution)[134,135], they will be treated independently, because they have a different impact on

the chromatographic data structure (relevant for the chemometric models discussed in Chapter 4). Unfortunately, it is not possible to identify a clear root-cause for each artifact, as they can result from the interplay of different factors, depending on the specific combination of sample, solute, and chromatographic conditions. The hope is that by describing the complexity of the chromatographic processes, the importance of considering these artifacts in a hypothesized data structure for a chromatographic dataset becomes visible.

For well-defined samples, as in a quality control laboratory, practical measures to account for many of the described artifacts could be implemented by sophisticated method optimization. The nature of untargeted analysis or semi-targeted analysis is, however, that little is known *a priori* about the sample composition. Hence, large concentration differences, undesirable matrix components, and a broad spectrum of chemically different analytes should be expected. Thus, artifacts are likely to be present in most untargeted datasets.

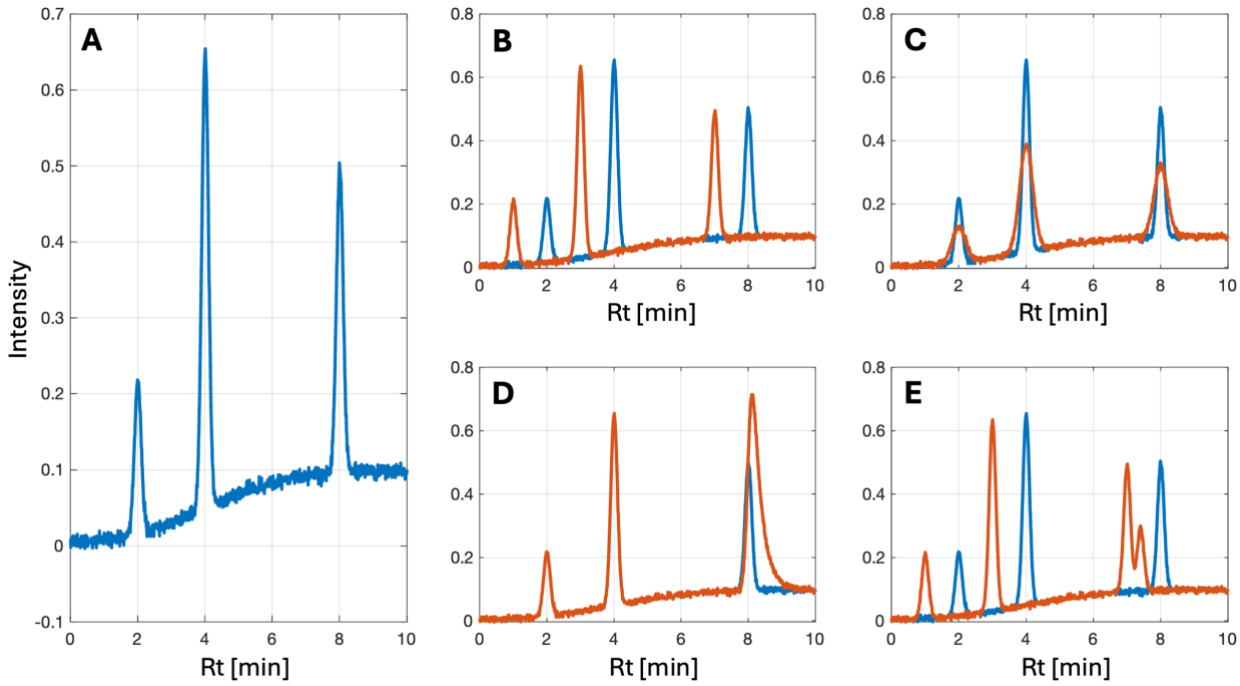


Figure 4*: Examples of gas chromatographic measurements showing different artifacts.* ***A:*** *Reference chromatogram.* ***B-E:*** *Reference chromatogram (blue) overlaid with chromatograms showing different artifacts (orange) These artifacts are referred to as: retention time shift* ***(B)****, band or peak broadening (****C****), peak skewness (****D****), and co-elution (****E****).*

3.2.1 What causes retention time shifts?

As mentioned in the beginning of this Section, the retention of a solute depends on many factors. A detailed understanding of the chromatographic process requires knowledge about the thermodynamic properties of the system and eventually a good model of the dynamics of the system using micro- and macroscopic kinetic models.[134,136]

The partition constant K_c expresses, at equilibrium state, how the solute is distributed between the mobile phase and the stationary phase. This relationship is described by Eq. 1, with c_s being the solute concentration in the solid phase, and c_m, being the concentration of the solute in the mobile phase.

Equation 1: $$K_c = \frac{c_s}{c_m}$$

From K_c we can derive the retention factor k, which is more commonly used in chromatography, under the assumption that the volume ratio β of the mobile phase and the stationary phase is constant (Eq. 2):

Equation 2: $$k = \frac{K_c}{\beta}$$

The larger k is, the more retained is the solute. If the dead time t_0 of the system has been experimentally determined, k can be used to calculate the retention time of the solute according to Eq. 3.

Equation 3: $$t_{R,s} = k * t_0 + t_0$$

The dead time is the time it takes for a compound that does not interact with the column to travel through the chromatographic system. Only the volume flow rate of the mobile phase and the length of the column define the dead time (under the strong assumption of constant pressure in the column). Hence, if in the example in **Figure 4B** the dead time and the retention times of the solutes are shifted linearly, it could be speculated that the retention time shift is due to instabilities in the volume flow rate or because the column has been trimmed (which is common practice in GC if "dirty" samples are measured).[137] Linear retention time shift refers to a shift in retention time of the same increment for all compounds. Retention times in GC and LC are

also sensitive to changes in the temperature, as can be seen by the approximation with the Gibbs-Helmoltz relationship shown in Eq. 4:[134]

Equation 4: $$\ln(k) = -\left(\frac{\Delta H^{\circ}}{RT}\right) + \left(\frac{\Delta S^{\circ}}{R}\right) + \ln(\beta)$$

Where ΔH° and ΔS° are standard enthalpy and entropy of the solvation process in the stationary phase, T being the absolute temperature and R the Avogadro constant. Eq. 4 is the two-parameter, ideal thermodynamical retention model which assumes that ΔH° and ΔS° are independent of the temperature, and the concentration.[129,134] Temperature increase leads to an exponential decay of k and shorter retention times. Additionally, in LC the dependence of k on the solvent strength is affecting the retention of a solute, which can for instance be described by the empirical model given by Eq. 5.[129]. In this model, c_0, c_1, and c_2 describe parameters that are fitted to the experimental data, and φ is the volume fraction of the organic solvent in a binary mixture. However, other empirical retention models have been suggested[138] and a more comprehensive overview can be found in literature.[139]

Equation 5: $$\log(k) = c_0 + c_1\varphi + c_2\varphi^2$$

Since k is a solute-specific quantity, temperature and even more the solvent gradient in LC will affect the retention of the individual solutes to a different extend. The same holds for changes of the properties of the stationary phase due to aging, the influence of pressure changes, and variations of pH in the mobile phase. A situation, in which the retention factors k_m are affected to a different extend by a change in the chromatographic conditions will be referred to as non-linear retention time shift, with $m \in \{1, \dots, M\}$ and M being the number of solutes.

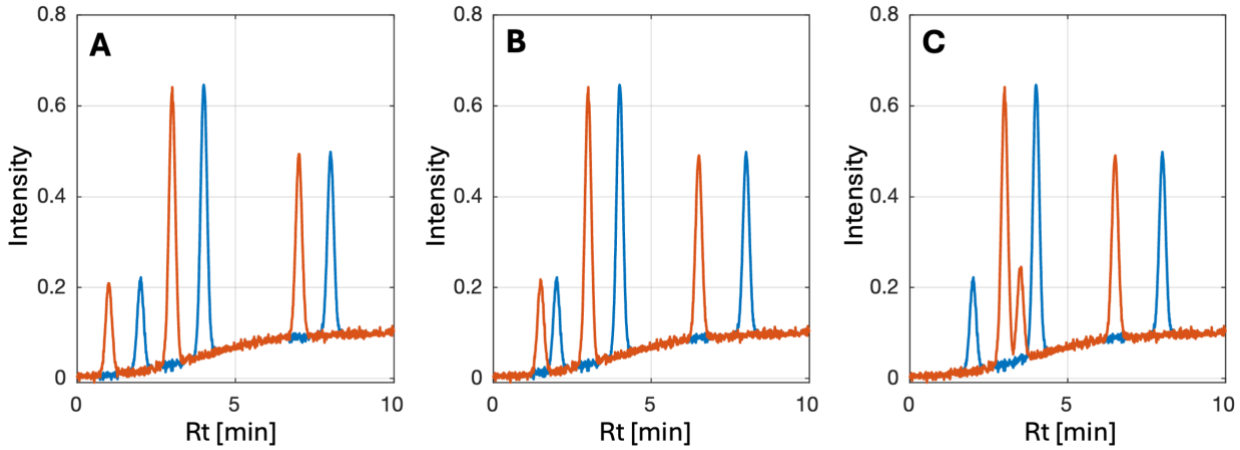


Figure 5*: Linear and non-linear retention time shifts.* ***A:*** *Linear retention time shifts, all solutes of the orange measurement are shifted by the same increment.* ***B****: Non-linear retention time shift, the increment of the shift increases for later eluting solutes.* ***C****: Non-linear retention time shift leading to a reversed retention order.*

Non-linear retention shifts are observed in GC and more frequently in LC.[140–142] Changes in the retention order have also been reported in LC, depending on the solvent gradient, and in GC depending on the heating rate.[134,143] **Figure 5** shows examples of linear (**Figure 5A**) and non-linear (**Figure 5B-C**) retention time shifts. For the discussion in Chapter 4, it is important to distinguish two cases: 1) retention time shifts that conserve the degree of overlap of neighboring peaks and their retention order (**Figure 5A-B**), and 2) retention time shifts that change the degree of overlap of neighboring peaks or the retention order (**Figure 5C**).

3.2.2 What causes peak broadening?

Peak broadening in chromatography can be explained by the competing effects of different diffusion mechanisms.[144,145] In principle, the previously described factors that are affecting retention of peaks are also affecting the peak broadening, as diffusion coefficients are for instance functions of temperature. For example, the Fuller-Schettler-Giddings model[146] describes the temperature-dependence of the binary gas diffusion coefficient D_{AB} with the empirically observed relationship $D_{AB} \propto T^{1.75}$. The diffusion processes that are often used to describe peak broadening are Eddy diffusion, longitudinal diffusion, and mass transport between the mobile and the stationary phase.[136,144] Eddy diffusion describes the peak broadening that happens when solutes travelling through the bed of a column take paths of different lengths to reach the end of the column. This effect can be more severe if channels in the packed bed are formed over time. In capillary columns used in GC, Eddy diffusion does not contribute to peak broadening. Instead, longitudinal diffusion and mass transfer are the primary

factors, and both depend on the diffusion coefficient of the solute in the mobile phase. A high diffusion coefficient leads to broader peaks due to increased longitudinal diffusion, while a low diffusion coefficient results in broader peaks due to slow mass transfer from the bulk to the stationary phase. Longitudinal diffusion is particularly significant in GC because diffusion coefficients of solutes in the gas phase are approximately 10,000 times larger than in a liquid.[134,136] Conversely, mass transfer limitations pose a more serious problem in LC. Mass transfer from the stationary phase to the mobile phase is affecting GC and LC and can be reduced by using a thinner film coating on the capillary or the solid support material.[134,136] Using smaller particle sizes for the column material in LC has the two positive effects: it reduces Eddy diffusion and increases mass transfer.[134,136]

Changes in the system variables (temperature, flow, pressure, mobile phase composition, pH), which affect diffusion coefficients, or other factors like the formation of channels in the chromatographic bed, can produce the artifacts as shown in **Figure 4C**.

3.2.3 What causes peak skewness?

Under ideal chromatographic conditions (also called linear chromatography), peaks have a Gaussian shape as can be predicted from diffusion models.[134,136] Deviations from this shape can be attributed to non-linear behavior of the partition constant K_c in regions of high solute concentrations. If the solute concentration is too high, the stationary phase gets saturated, which means that a fraction of the solute molecules will not interact with the stationary phase. Consequentially, some solute molecules will travel faster through the column because they are less retained. This situation is called "fronting" and can be modeled by the Langmuir adsorption isotherm.[147]

Figure 4D shows the reversed situation, where the peak is right skewed. The illustrated phenomenon is called "tailing". Tailing can occur, when competing retention mechanisms exist, for example when polar solutes interact with free silanol groups of the column material.[148,149] Since these polar-polar interactions are stronger than the interaction of the same polar solute with the non-polar stationary phase coating, the former described interaction leads to stronger retention of the solute. There are usually few active silanol groups and therefore only a fraction of the solute is retained stronger by this secondary retention mechanism, which leads to the asymmetric peak shape. The activity of the silanol groups can be controlled by adjusting the pH and by adding buffer salts.[148,149] This, however, can lead to problems in combination with mass spectrometry detectors, which will be discussed later.[150,151] Furthermore, a mismatch between the sample solvent and the mobile phase may result in non-equilibrium conditions due

to viscous fingering[152], which will lead to low reproducibility of peak shapes, even under otherwise tightly controlled chromatographic conditions.[153] Solvent incompatibility is one of the big challenges in comprehensive LC×LC. This is, because two columns with orthogonal retention mechanism are connected to maximize the separation performance.[154] Hence, the solvent strength and viscosity of the first retention dimension is not matching the solvent strength and viscosity of the second dimension and viscous fingering can occur when the first-dimension fractions are injected into the second column.[154] Recent developments in modulation technique try to overcome this problem.[155–157]

3.2.4 Co-elution and deconvolution

The separation of two solutes under specific chromatographic conditions requires that their k values differ. Differences in k values can be translated into differences in the retention time $t_{R,s}$, according to Eq. 3. The retention time difference Δt_R required for two solutes to be fully resolved depends on their peak broadness σ, which is affected by the factors described in Sections 3.2.2 and 3.2.3. Under the assumption of Gaussian peaks, all having the same variance σ^2, the minimum retention time difference $\Delta t_{R,min}$ for two solutes eluting at $t_{R,1}$ and $t_{R,2}$ can be calculated according to Eq. 6.[134]

Equation 6: $$\Delta t_{R,min} = \Delta s_{min} \sigma$$

In Eq. 6, Δs_{min} is a factor that determines how many standard deviations should separate two neighboring peaks, with a recommended minimum of four to achieve an acceptable peak resolution.[108] However, the sufficient value of Δs_{min} also depends on the height difference between the peaks; for example, larger values of Δs_{min} may be required to fully resolve a small peak eluting next to a large peak.

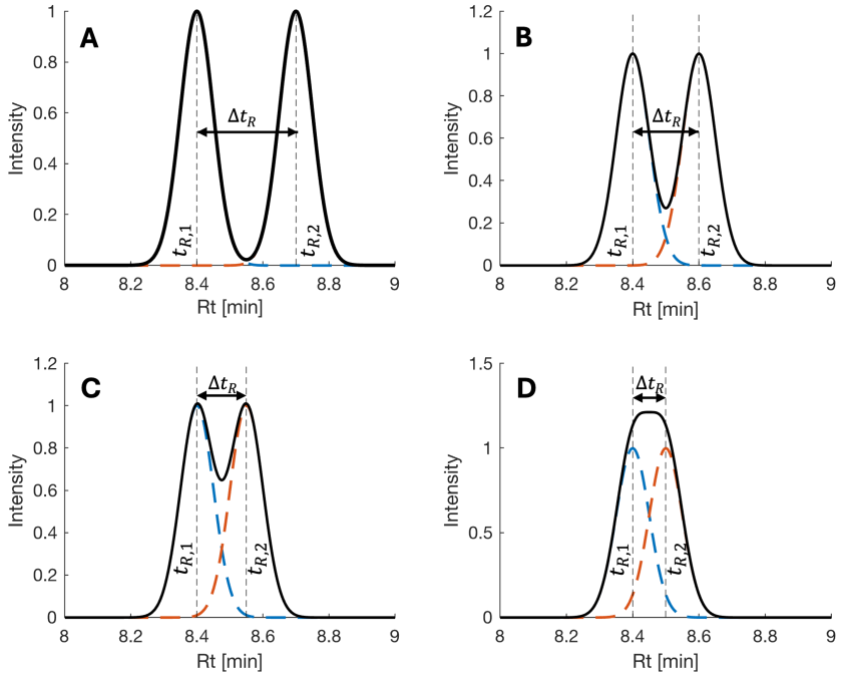


***Figure 6**: Two neighboring peaks with different degrees of co-elution. The degree of co-elution has been modified by choosing different values for Δs (see Eq. 6). Specifically, Δs values have been set to six (**A**), four (**B**), three (**C**), and two (**D**). While peaks in **A** and **B** show acceptable resolution, the overlap of the two peaks in **C** and **D** is so severe that a reliable quantification would be difficult in practice. If peaks have different heights, even larger values of Δs are required for sufficient resolution.*

Figures 6A-D show examples of two peaks of identical height that are resolved to different degrees, with Δs varying between six in **Figure 6A** to two in **Figure 6D**. Even at Δs values of four the peaks are considerably overlapped. In **Figure 6C** the two peaks are severely overlapped for an Δs value of three and in **Figure 6D** the two different peaks cannot be distinguished at all for an Δs value of two.

Co-elution are problematic because they can lead to inaccurate determination of the peak area, which will cause errors in the quantification of the respective analytes in the sample. Further, it may happen that individual analytes are not even detected, if the co-elution is as severe as in **Figure 6D** or if the intensity differences between the two neighboring peaks are larger. If it is assumed that all solutes are equally spaced along the retention time, the number of peaks that can maximally be resolved with a given chromatographic method can be calculated according to Eq. 7:

Equation 7: $$N_p = \frac{t_G}{\Delta t_{R,min}}$$

With t_G representing the duration of the chromatographic method and $\Delta t_{R,min}$ being the minimum retention time difference between two neighboring peaks (as defined in Eq. 6), N_p is

defined as the peak capacity. The assumptions that have been made above are very optimistic and not found in practical applications. However, to give a numerical example, let $t_G = 60$ min, $\sigma = 0.05$ min, $\Delta s_{min} = 4$, then $\Delta t_{R,min} = 0.2$, according to Eq. 6, and $N_p = 300$, according to Eq. 7. In a recent study it has been estimated that peak capacities of about 300 are the maximum that can be achieved for a common LC separation problem.[158] This number may seem high for targeted analysis, if the objective is to analyze mixtures of 10-20 analytes but considering that samples in untargeted metabolomics can contain more than 1000 analytes[93], this number can be discouraging. Even more so, if it is considered that results from *Statistical Overlap Theory* predict that peak capacities need to be a factor of ~20 larger to ensure a 90 % probability of resolving all solutes sufficiently.[135] This means, that statistically a peak capacity of 300 is only sufficient to resolve up to 15 analytes in a sample. One obvious way of improving this situation is to increase the peak capacity of chromatographic systems dramatically, which can be achieved with comprehensive two-dimensional chromatography.[159,160] Another option is to use selective multi-channel detectors, which provide a solute-specific response. The mass spectrometry detectors, discussed in the next Chapter, belong to this category. The third option is to use computational tools to resolve overlapped signals and extract chemical information mathematically (discussed in detail in Chapter 4 and Chapter 5). For the characterization of complex samples, it may be required to employ all three options together.[23,161–163]

The use of computational tools provides some additional advantages, which are illustrated using a simulated, complex chromatographic region from a GC-MS dataset as an example. In this dataset four solutes of different concentrations are co-eluting. **Figure 7A** shows five measurements in which the four solutes elute broadly in the same chromatographic region. The colored lines show the peaks of the individual solutes, and the black dashed line shows the sum of all peaks (total ion-chromatogram, abbreviated to TIC). The peaks have been subjected to non-linear retention time shifts and their concentration varies in the different samples. On the other hand, the TIC appears as a single peak where the shape of the peak is changing across the five sample measurements. The overlayed mass spectra of the individual solutes are largely overlapping which is a realistic situation in GC-MS (e.g., **Figure 7B,** m/z = 100). **Figure 7C** shows a 2D plot of the mass detector signal recorded for the first sample. Although it might be possible to manually find selective mass channels to isolate the different peaks, it requires a lot of manual work. Instead, the co-eluting signals can be perfectly resolved using the SIT algorithm [Paper I]. The results of the deconvolution can be seen in **Figure 7D-E**, in which the modeled elution profiles and mass spectra are depicted as blue lines and are compared to the true elution profiles and mass spectra depicted as orange lines (mass spectra shown as mirror

plot). The modeled elution profiles and mass spectra perfectly match the true elution profiles and mass spectra, and it is even possible to separate the baseline signal from the solute signals (last row in **Figure 7D-E**). Hence, the peak areas of the modeled peaks can be used as relative concentrations without further pre-processing steps, such as baseline removal. As a result, computational tools have the potential to greatly facilitate the extraction of accurate chemical information from complex chromatographic datasets. As previously stated, it is important that the model assumptions are matching the data structure of the chromatographic data in good approximation (detailed in Chapter 4).

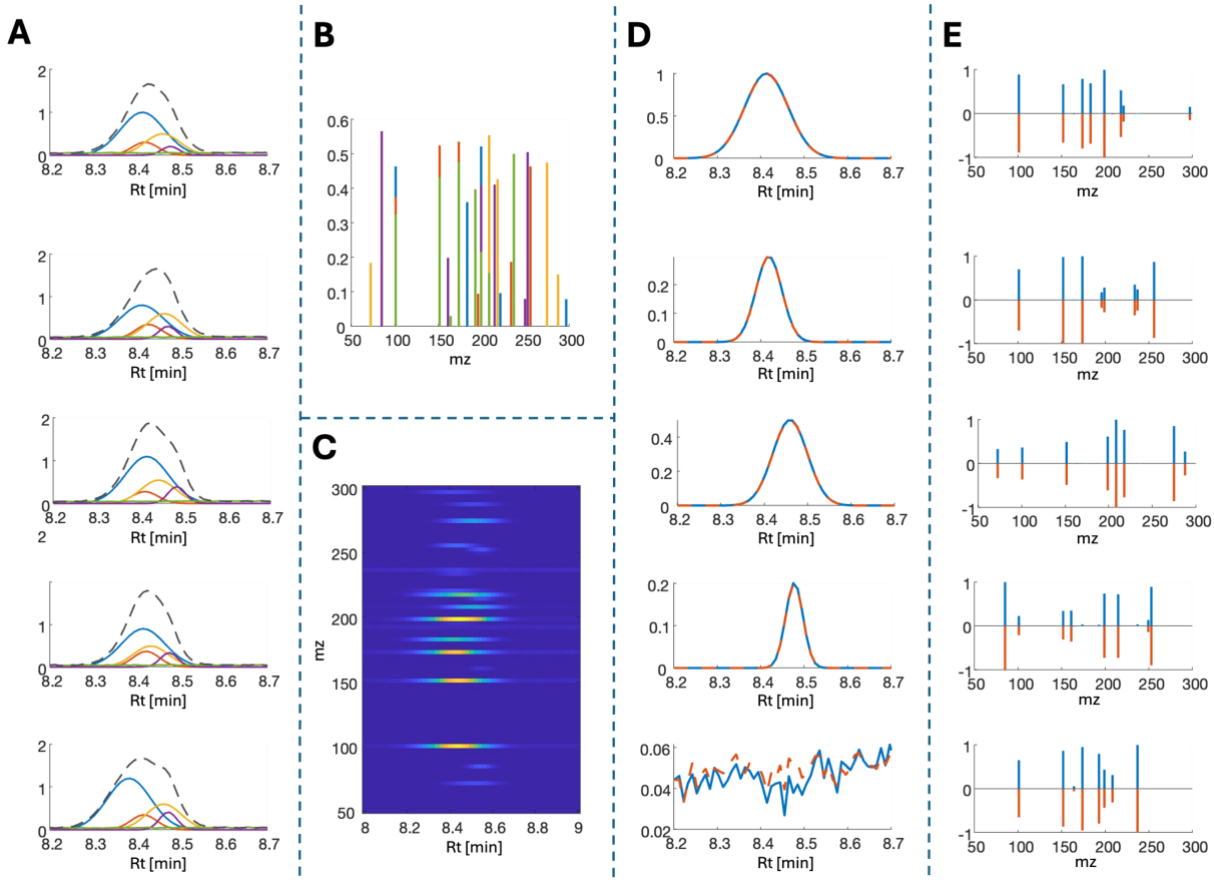


***Figure 7**: The benefit of using computational tools for the deconvolution of complex chromatographic data is illustrated. **A**: Dataset consisting of five samples with four co-eluting solutes (colored peaks). The TIC is shown as black-dashed line. **B**: Overlaid mass spectra of the solutes. **C**: 2D detector signal of the first sample. **D-E**: Results of the deconvolution. As example deconvoluted elution profiles and mass spectra of the first sample are shown. Blue lines indicate the modeled elution profiles and spectra and orange lines the true signals.*

3.3 History of mass spectral detectors hyphenated to chromatography

In **Figure 2**, a mass spectrometer was referred to as a detector, which can be interpreted as the "chromatographers view" on mass spectrometry. In opposition, a mass spectrometry expert may see a chromatographic instrument just as another form of sample inlet. Although the importance

of mass spectrometry as an individual instrumental technique cannot be held high enough, this work will be biased slightly toward the chromatographer's view. Nevertheless, to give the due credit the following short history will present some of the most fundamental breakthroughs enabling the use of mass spectrometry coupled to chromatography instruments.

In 1956, Roland S. Gohlke and Fred McLafferty presented their work on the first GC-TOFMS.[164] This breakthrough was based on prior work, including the development of the electron ionization (EI) method by Bleakney and Nier[165,166], as well as the TOF mass analyzer by Stephens.[167] In the following years, researchers discovered and developed alternative ionization methods that would allow for softer ionization compared to EI. A softer ionization method means that molecules are ionized under less harsh conditions (lower ionization energy) which yields higher fractions of unfragmented, molecular ions. Examples of these soft ionization techniques are field ionization (FI) and chemical ionization (CI).[168,169] A recent review of untargeted analysis studies in environmental sciences summarized that EI is currently by far the most popular ionization technique for GC-MS.[101] This is likely due to the high reproducibility and standardization of EI mass spectra measured with 70 eV ionization energy.[170] The development of LC-MS instruments turned out to be difficult due to the "volatility barrier" and required the development of ion sources that could operate under atmospheric pressure conditions.[171] The first prototypes of LC-MS instruments were proposed by Horning and co-workers in the mid-seventies, based on atmospheric pressure chemical ionization (APCI).[172,173] However, a big breakthrough happened in 1988 when John Fenn published the first successful LC-MS coupling using the electron spray ionization (ESI) technique[174], which had been previously invented in 1968 by Malcolm Dole.[175] Briefly, ESI works by applying a high voltage to a liquid sample, creating a fine spray of charged droplets. Coulombic repulsion causes these droplets to break apart, allowing the solvent to evaporate under atmospheric conditions, and the analyte molecules to form gas-phase ions for mass spectrometric analysis. The broad application range of ESI (considering polarity and size of analytes) as compared to EI and APCI, for instance, make it the most widely used ionization technique for LC-MS today.[101,176] Additionally, the technological improvements of mass analyzers led to an increasingly high resolution of mass spectrometers. Some of the key developments were the time-of-flight analyzer[167], quadrupole analyzer[177,178], cyclotron resonance analyzer[179,180], and the orbitrap analyzer[181]. Hence, LC-MS and LC-MS/MS are relatively young analytical instruments compared to GC-MS, which is one of the reasons why standardization has not yet reached the same level as in GC-MS.[170,182] **Figure 8** provides examples of the essential parts of a mass spectrometer, showing examples of ion sources

(**Figure 8A**), mass analyzers (**Figure 8B**), and detectors (**Figure 8C**). The discussion in the following Sections will mainly be restricted to the instrumentation examples given in **Figure 8.**

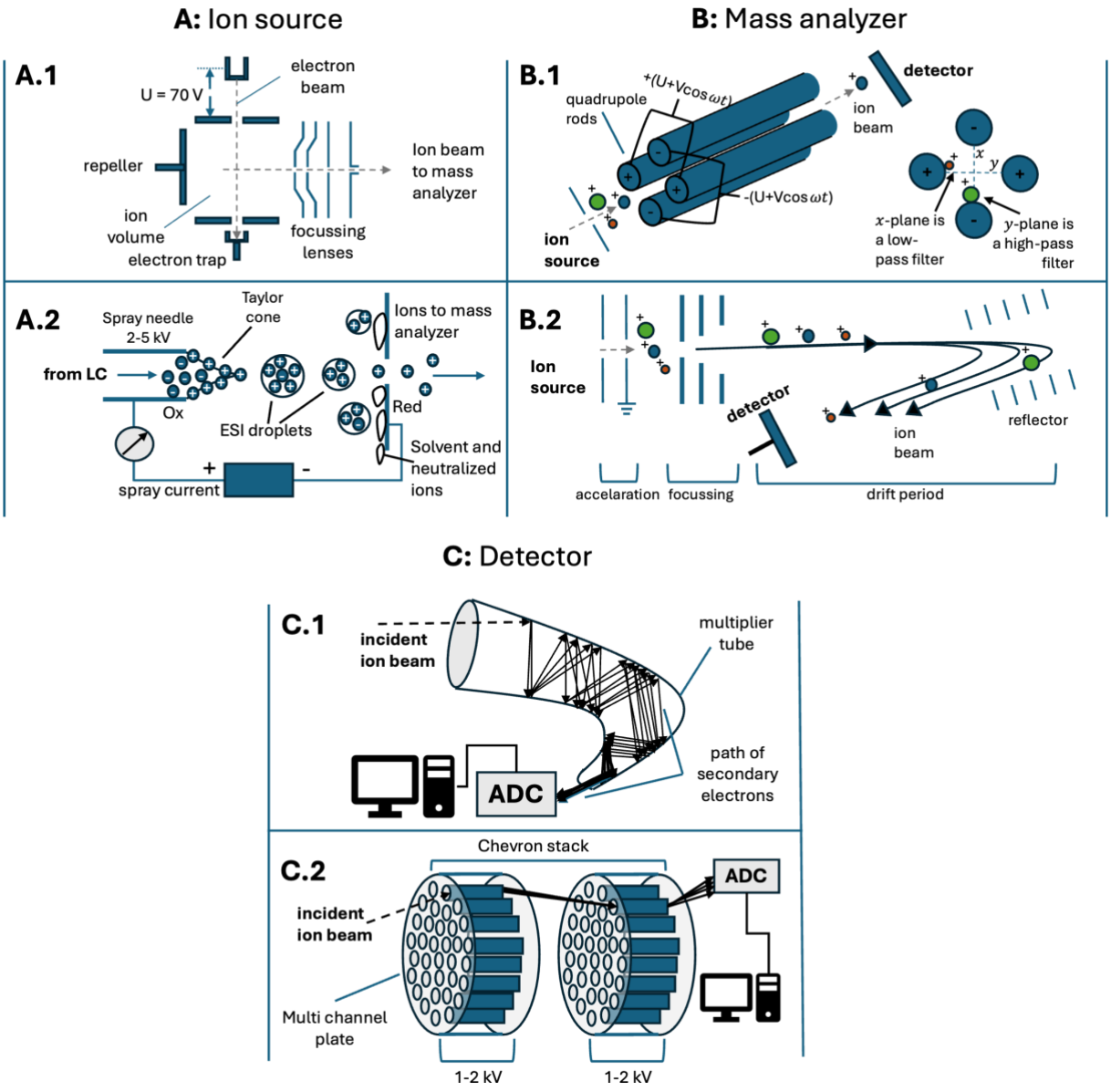


Figure 8*: Visualization of the essential components of a mass spectrometer, inspired by Gross*[183] */ Chec and Enke*[184]*. **A**: Examples of ion sources, **A.1** Election ionization (EI), **A.2** Electron spray ionization (ESI). **B**: Examples of mass analyzer, **B.1** Quadropole mass analyzer, **B.2** Time-of-flight mass analyzer, **C**: Examples of detectors, **C.1** Channeltron photomultiplier, **C.2** Multi-channel plate (MCP)*

3.4 Mass spectral acquisition modes and artifacts

In a mass detector, the solutes in the effluent from the chromatographic system are ionized (either under vacuum or atmospheric pressure conditions), the created ions are analyzed and a signal proportional to the abundance of the specific ions is recorded over time. If a soft ionization source like CI[185] or ESI[176] is used, a larger fraction of the ions is composed of the molecular ion of the solute and conversely more characteristic fragments are produced, if a harder ionization like EI[186] is used.[187] Depending on the resolution of the mass analyzer, it is possible to measure the mass-to-charge ratio (m/z or mz) in lower or higher accuracy. The mass spectral data will be classified as low resolution ($\frac{m}{\Delta m_{50\%FWHM}}$ < 10.000) or high resolution ($\frac{m}{\Delta m_{50\%FWHM}}$ > 10.000), with $\Delta m_{50\%FWHM}$ being the width of a mass peak at 50 % of its maximum intensity, and m being the ion mass at peak apex.[188]

With respect to the construction of the recorded signal, two sampling steps need to be considered. The first sampling step is the sampling of the chromatographic time-domain and the second is the analog-to-digital conversion of the signal recorded at the detector.[183] The second step will be discussed in more detail in Section 3.5. The sampling frequency of the time-domain, expressed as scans per second, depends on the total cycle time of the mass detector. The cycle time of a mass analyzer is defined as the sum of all instrument tasks required to finish one scan (which includes analog-to-digital conversion).[189] The duty cycle of a mass analyzer, defined according to Equation 7, is often used to assess its efficiency.[183,189]

Equation 7: $$Duty\ Cycle = \frac{Time\ to\ analyze\ the\ ions}{Time\ to\ finish\ a\ scan}$$

The cycle time varies between different analyzers and is also dependent on the mass spectral acquisition mode and the required mass accuracy.[183,189,190] A higher number of scans and a higher accuracy will yield a longer cycle time and a lower sampling frequency.[183,189,190] Additionally, a high number of scans may also be required to improve the signal-to-noise ratio (SNR).[183,189,190] Generally, scanning mass analyzers like the quadrupole yield lower duty cycles compared to TOF mass analyzers.[183,190,191]

Figure 9 shows simulated data examples of chromatography hyphenated mass spectrometry data measured in different acquisition modes. **Figure 9A** shows a full scan measurement in combination with a hard ionization, **Figure 9B** shows a selected ion measurement (SIM)[192], **Figure 9C** shows a MS measurement acquired in full scan mode with soft ionization[193], **Figure 9D** shows the MS^2 measurement of the same measurement acquired in data dependent

acquisition (DDA)[194], and **Figure 9E** shows the MS^2 (also called MS^E)[195] measurement in data independent acquisition (DIA) all ion fragmentation.[196] The full scan mode shown in **Figure 9A** resembles a GC-MS measurement, which produces many fragments. Two compounds are co-eluting (blue and orange signals), and an additional background signal is present (grey). If the goal is to quantify one targeted compound, e.g., the blue-colored peak, the signals from the co-eluting compound and the background might be disturbing. Hence, it is more feasible to only record specific fragments that are unique to the target compound, which is done in SIM (**Figure 9B**). The additional advantage is that the SNR will be improved (assuming a constant sampling frequency), if only a few fragments are scanned repeatedly, instead of scanning the full m/z range. In **Figure 9C**, only the molecular ion peaks of three compounds are shown, resembling a soft ionization (in-source fragments, isotopic fragments and adducts are not considered, to aid visualization). Since two of the compounds have the same molecular mass (up to the mass accuracy of the mass analyzer), a characterization is not possible without further information. Tandem mass spectrometry can be used to obtain additional fragmentation information. In DDA, only ions of a specific m/z (called precursor or parent ions) are fragmented into product ions. To analyze the fragment ions, either the full mass range can be scanned to obtain the complete fragmentation spectrum or only characteristic fragments of the target compound are selected.[197] However, like in SIM, DDA mode has the drawback, that only targeted compounds will be found.[193] Conversely, in DIA all ion fragmentation, full fragment spectra are generated for all precursor ions. Similar to the full scan GC-MS measurement, the DIA data is very complex and deconvolution (see **Figure 7**) may be required to extract clean fragmentation spectra that can be compared to a database.[34,193,194,196,198] **Figure 9A** and **Figure 9E** resemble acquisition modes typical for an untargeted analysis approach[198,199], while **Figure 9B** and **Figure 9D** resemble acquisition modes commonly used for targeted analysis.[197,200] In terms of instrumentation, the most common analyzer for GC-MS is the quadrupole, whereas TOF mass analyzers are almost exclusively are used for GC×GC-MS, due to their fast acquisition time.[183,190,201] Tandem mass spectrometry requires the combination of two or more mass analyzers. Usually, triple quadrupole analyzers are used for DDA experiments like multiple-reaction monitoring, while the combination of quadrupole and time-of-flight analyzer (QTOF) can for instance be used for DIA experiments.[196,197]

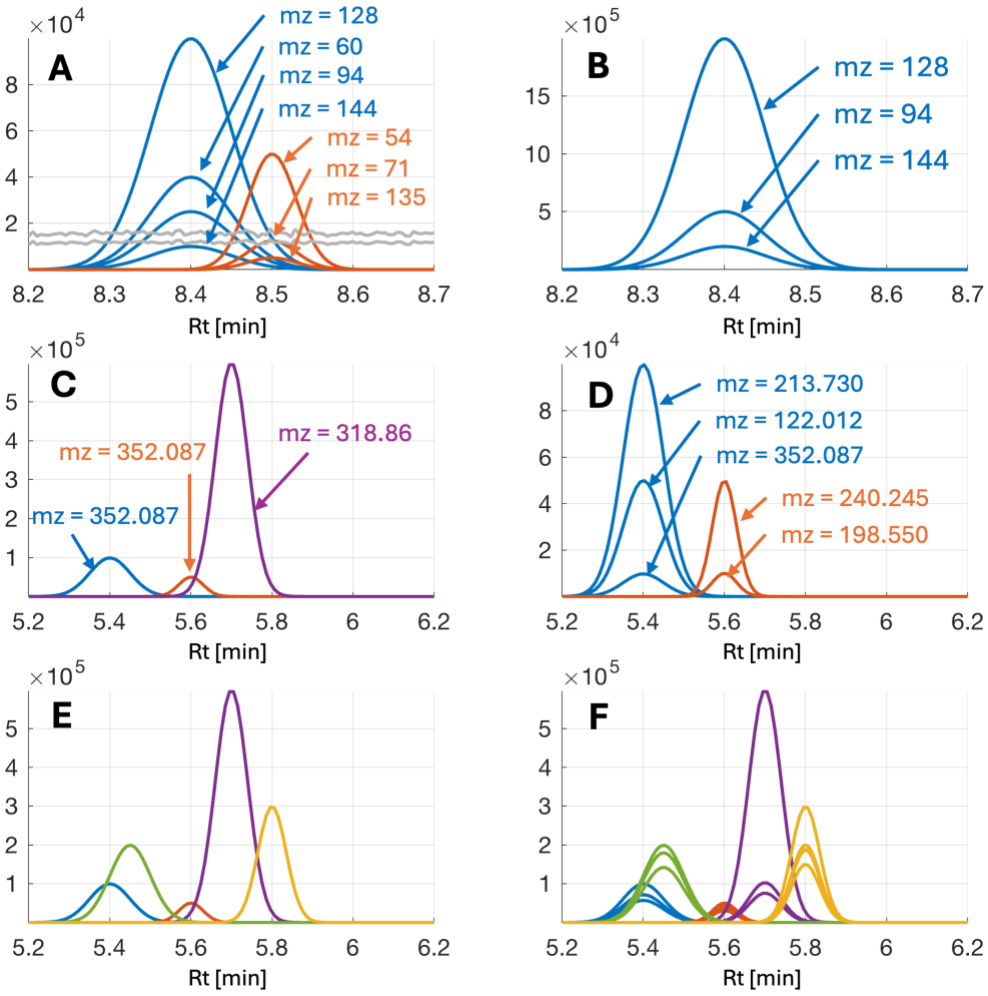


***Figure 9**: Visualization of different acquisition modes. **A**: Full scan acquisition of low-resolution GC-MS, **B**: Selected ion mode (SIM) of the same sample shown in (**A**). **C**: Full scan MS^1 measurement of data-dependent acquisition (DDA) experiment, showing the precursor ions. **D**: Selected precursor ions are fragmented, and specific fragments / transitions are monitored in MS^2 (e.g., multiple reaction monitoring). **E**: Full scan MS^1 measurement of data independent acquisition (DIA). **F**: All precursor ions from MS^1 are fragmented yielding MS^2 fragment spectra for all compounds.*

Targeted studies usually make use of reference standards for identification and quantification of the targeted compounds which is economically not feasible for untargeted studies due to the large number of compounds.[54,68,71] Given the high specificity of mass spectra (especially in high-resolution), structure assignment by comparison of experimentally measured mass spectra against mass spectral databases is a viable approach.[68,71] However, two prerequisites for this approach are 1) sufficiently large databases and 2) a sufficiently high reproducibility of the mass spectra measured in different sample matrices, on different instrumental platforms and across multiple laboratories.[182,202] This aspect is further discussed in Section 5.1.4. In the following Sections 3.4.1, 3.4.2, and 3.4.3, the focus will narrow to investigate which factors can affect spectral reproducibility within a set of samples measured on the same instrument, in

the same laboratory. This can be considered the prerequisite of a prerequisite and is fundamentally important for the discussion of the feasibility of chemometric models in Chapter 4. Chemometric models assume that mass spectra recorded in different scans across a chromatographic peak are consistent and further, that the mass spectra of a given analyte are highly reproducible across multiple samples in a series measurement. In the following Sections, a few examples will be discussed for which both assumptions do not hold.

3.4.1 Spectral skewing

Scanning mass analyzers like a quadrupole acquire mass spectra by measuring the abundance of ions one after another, by scanning through a pre-selected mass range.[183,201] Conversely, a TOF analyzer measures simultaneously the abundance of all ions in one pulse (see **Figure 8**).[183] Consequently, scanning mass analyzers require a longer time to collect data over a larger mass range. This can lead to problems with peak resolution, if the chromatographic peak width is small relative to the total cycle time of the quadrupole. However, small peak widths are generally desirable because they provide higher peak capacities and allow for shorter chromatographic run times (see discussion in Section 3.2.4).[16,159] In GC×GC applications, for instance, peak width can be as small as 60 ms[191], which under the assumption of a minimum number of 12 scans per peak, requires a scan frequency of 200 Hz. Some studies have shown that quadrupole analyzers can operate with up to 50 Hz for a limited scanning range of 290 amu[203] (scanned mass range) and demonstrated successful applications in GC×GC[203,204]. However, TOF is the predominantly used mass analyzer for GC×GC.[160] Conversely, quadrupole is the most abundantly used mass analyzer for one-dimensional gas chromatography.[201]

The low sampling frequency of the quadrupole analyzer can cause distorted chromatographic peak shapes, leading to problems in quantification.[205] A related issue is spectral skewing, which is caused by concentration changes in the ion chamber during the acquisition of a scan.[201] Spectral skewing is illustrated in **Figure 10**. In **Figure 10A** a simulated Gaussian profile has been sampled with 16 (blue) and with four (orange) data points to create two TIC profiles. The interpolated blue TIC profile is perfectly matching the original Gaussian profile, while the interpolated orange TIC profile shows larger deviations. In **Figure 10B**, the problem of spectral skewing is exemplified using the normalized TIC profile constructed with a lower sampling frequency. Since the analyte concentration is changing during the duration of a scan, ions with high m/z have a higher relative abundance (if the concentration is increasing) or a lower relative abundance (if the concentration is decreasing), compared to ions with lower m/z. This effect is

of course exacerbated if the concentration change between two scans is large, which is the case when a peak is only sampled with few scans. In **Figure 10C-D** the effect on the extracted ion chromatograms (EICs) of different m/z is shown. Only small differences in the shapes of EICs recorded at different m/z are visible for the peak sampled with 16 scans. However, larger differences between the shapes of the EICs can be seen for the peak sampled with four scans. The difference between low and high sampling frequencies is indicated by a more significant shift in the peak maxima at low sampling frequency, shown by the vertical black dashed lines. Hence, the consistency of mass spectra across scans depends on the sampling frequency if a quadrupole mass analyzer is used. This should be considered as a factor influencing the chromatographic data structure, for instance in fast GC-MS methods.[205] Smoothing methods can eventually help to reduce the effect of peak skewing (see Supporting information of [Paper IV]).

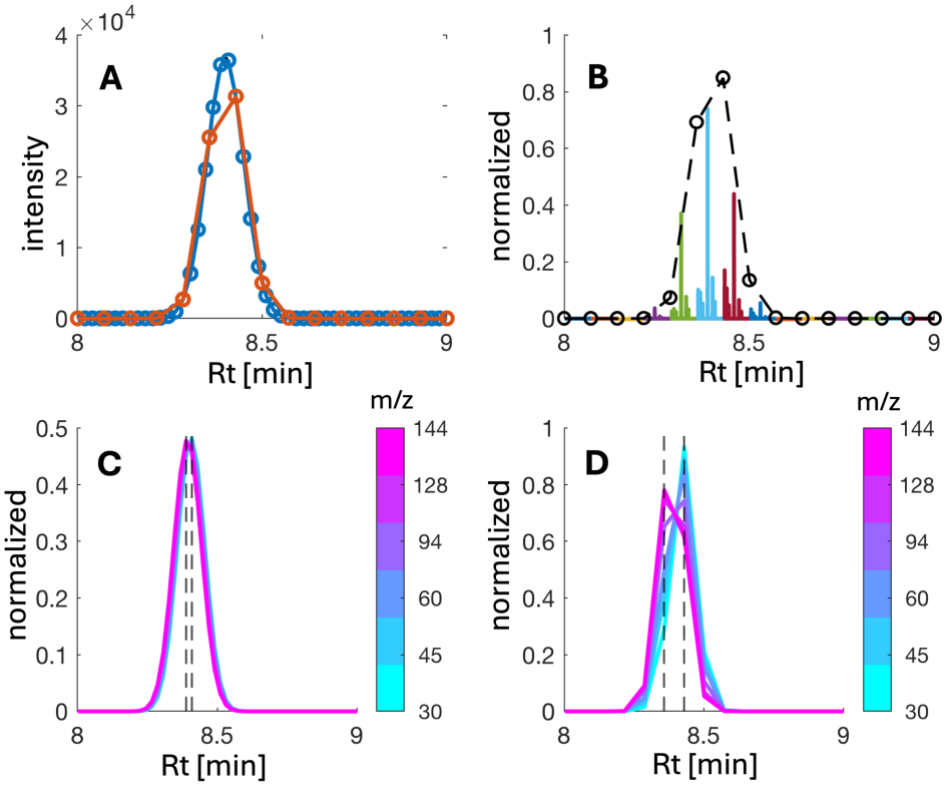


***Figure 10**: Visualized is the effect of spectral skewing. **A**: TIC profiles created by sampling a Gaussian elution profile with 16 (blue) and 4 scans (orange). **B**: Illustrating the mechanism causing spectral skewing using the TIC profile sampled with 4 scans. The concentration of the ions changes during the duration of a scan. **C-D**: Effect of the spectral skewing on the shape of the EICs. Different colors indicate different m/z values according to the color bar. While the shapes of the EICs are only mildly affected in the peak sampled with 16 scans, larger differences can be seen between the EICs of the peak sampled with 4 scans.*

3.4.2 Saturation effects

It has been discussed in Section 3.2.3, that too high solute concentrations can cause non-linear behavior of the chromatographic separation process resulting in skewed chromatographic profiles. Moreover, high concentrations can also lead to saturation of the mass detector which can cause peak distortion of the EICs of highly abundant ions.[206,207] Saturation of the mass detector essentially means that the detector is not recording a response that is linearly proportional to the number of solute molecules eluting from the chromatography column. Thus, saturation effects may lead to inaccurate quantification of analyte concentrations and have an impact on the data structure. More specifically, saturation can be attributed to charge-competition leading to insufficient ionization (especially in ESI)[207,208], or to saturation of the analog-to-digital converter (ADC)[209,210] / time-to-digital converter (TDC)[206,211].

Hardware improvements over the past decades have extended the dynamic range of quadrupole and TOF analyzers considerably.[209,211,212] The dynamic range refers to the concentration range over which the detector provides a linear response. Furthermore, different strategies for instrumental tuning have been suggested to limit the effect of saturation.[207] Switching to a different ion source or sample dilution are alternative approaches to deal with saturation stemming from insufficient ionization.[213,214] Another approach, mainly applicable to saturation of the ADC / TDC, is the algorithmic correction of saturation effects. This strategy comprehends for instance the correction of detector dead time (not to be confused with the dead time t_0 of a chromatographic method), or post acquisition peak shape modelling using unsaturated EICs.[206,210,215] Finding unsaturated EICs for example from the isotopic envelope is more straightforward when a soft-ionization technique such as LC-ESI-MS has been used.[210] However, for low-resolution GC-EI-MS the procedure suggested by Bilbao et al is difficult to implement due to in source fragmentation and the resulting low abundance of the molecular ion and its isotope peaks.[210]

Saturation effects remain a challenge for untargeted studies, such as in metabolomics, where analyte concentrations can vary over up to eight orders of magnitude, exceeding the dynamic range of current mass detectors.[206] If saturation effects are not corrected, they will lead to biased mass spectra and poor reproducibility of mass spectra across samples, besides the already mentioned problems in quantification. In **Figure 11**, saturated signals of a GC-MS measurement (**Figure 11A**) and ^{2}D elution profiles from a GC×GC-TOFMS (**Figure 11B**) are shown. The data shown in **Figure 11A** comes from an unpublished study in which volatiles formed during apple-wine fermentation have been analyzed. The data shown in **Figure 11B** represents the highest calibration point of the calibration data described by Armstrong et al.[216]

In the GC-MS data, saturation effects can easily be spotted from comparing the individual EICs. In the case that the ion count of a given m/z exceeds a certain threshold value, the intensities get cut. This results in a flat top of the saturated EIC (orange line in **Figure 11A**). Hence, EICs with intensities below the detector cut-off show no saturation effect, have a linear detector response (blue lines in **Figure 11A**), and can be used to model the peak shape of the saturated EIC. Conversely, in **Figure 11B** all EICs (blue- and orange-colored lines) have a non-linear detector response, which has been confirmed by comparing the EIC peak areas to the peak areas of lower concentrated calibration points. However, only some of the EICs show clearly distorted peak shapes. The distortion does not happen at a specific ion count threshold, such that some EICs of highly abundant ions have reasonable peak shapes (blue-colored) whereas EICs of lower abundant ions show a flat top (orange-colored). This observation has also been replicated for some of the higher concentrated calibration standards in the GC×GC-TOFMS reference dataset published by Weggler et al..[217] Unfortunately, the literature on detector saturation of TOF mass analyzers coupled to GC or GC×GC is not very comprehensive.[206] Hence, without further information, it seems to be more difficult to find a correction procedure for this type of data. For completeness, a blog post from the Sciex corporation provides a rule of thumb for how to calculate intensity thresholds of EICs for their LC-TripleTOF 5600+ and 6600+ instrument series.[218]

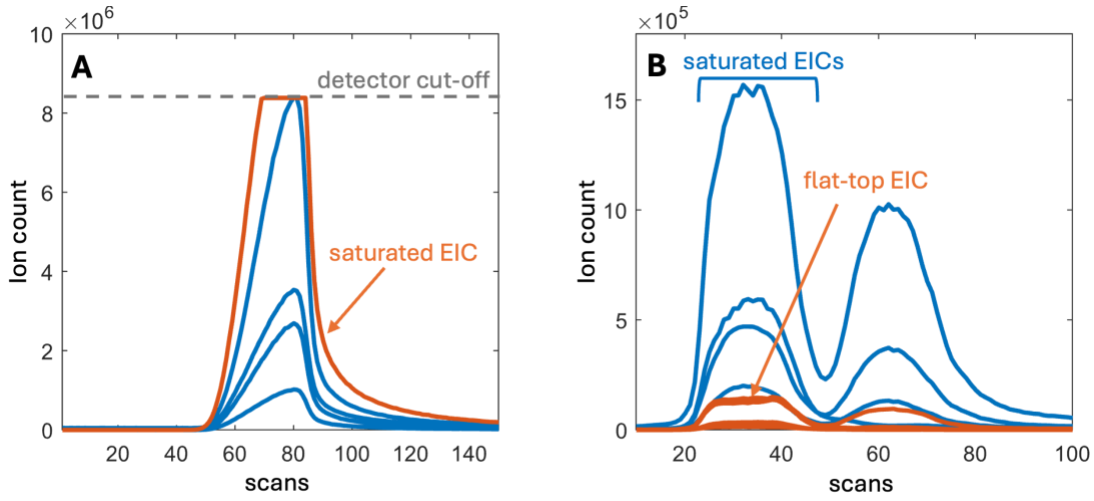


***Figure 11**:Comparison of detector saturation on different mass analyzers. **A**: Different EICs measured on a GC-MS equipped with a quadrupole mass analyzer. The ion count of the orange-colored EIC reaches the detector cut-off, resulting in peak saturation where the peak top is cut off. The blue-colored EICs, however, do not show saturation. **B**: Different ^{2}D-EICs from a GC×GC-TOFMS measurement. Although all EICs exhibit a non-linear detector response (as discussed in the main text), only the orange-colored EICs display clear signs of detector saturation, characterized by a flat peak top.*

3.4.3 Ion suppression and adduct formation

The development of atmospheric pressure ionization and especially the development of the ESI source was the technological breakthrough that enabled the widespread use of LC-MS(/MS), as mentioned in Section 3.3. To date, ESI remains the most widely used ion source in LC-MS(/MS). However, a significant drawback of the ESI source is that its ionization efficiency is highly dependent on the sample matrix and is strongly influenced by co-eluting compounds.[151,219] The presence of sample matrix components, co-eluting analytes, ion-pairing agents, and modifiers used to improve chromatographic separation can either suppress or enhance the formation of molecular ions.[150,151,220] These ion suppression or enhancement effects are specific to the molecular properties of the analyte.[220,221] Assessing the impact of matrix effects and co-elution on ionization efficiency is crucial for accurate quantification in targeted studies and for comparing relative concentrations across samples in untargeted studies.[71,76] Two experimental strategies that are often employed are post extraction spiking and post column infusion of internal standards that can be used for correction.[219,222] It has been pointed out that conditions typically found in untargeted studies (large number of co-eluting peaks, minimal sample preparation) will worsen the impact of matrix effects.[221] A more recent study suggested an adapted procedure of the post column infusion method together with quantitative structure relationship modelling for the correction of matrix effects in untargeted studies.[223]

A consequence of matrix effects that requires special consideration is the potential change in the relative ion abundances in an analyte's mass spectrum across samples, for instance due to adduct formation.[150,184,221] Adducts are typically formed with inorganic or organic cations such as sodium, potassium, or ammonium, but also solvent adducts with water or acetonitrile are frequently observed.[150,184] As a result, analyte mass spectra measured on LC-MS(/MS) instruments under varying matrix effects cannot be expected to be consistent across samples.[198]

3.4.4 Raw data pre-processing

The detector of a mass spectrometer records mass spectra at different retention time points as digital signals by sampling the continuous analog signal generated by a photo multiplier, using an analog-to-digital converter (see **Figure 8**). In modern instruments, the digitized signal is recorded in profile mode and stored as a discretized-continuous signal[224], whereas older instruments employed on-the-fly compression to reduce the size of the stored data.[225] However, to compress the enormous data size of high-resolution instruments, profile-mode mass spectra are usually still transformed post-acquisition, into centroided mass spectra, before data analysis.[225,226] In centroided mass spectra, only the apex of a detected mass peak is stored

(intensity and m/z value), which allows keeping the data in more efficient, sparse data formats.[227] The conversion from profile-mode mass spectra to centroided mass spectra, however, results in the loss of information about the mass peak shape, which could be valuable for diagnostics.[225,228] Although it is acknowledged that profile data is richer in information[225,228], Vereyken et al. showed that centroiding can be beneficial for qualitative analysis.[224] The vast majority of existing data analysis methods and workflows, including those discussed in Chapter 4 and 5, are capable of processing (or have only been applied to) centroided data. [35,37,84–89] Centroiding can either be performed in vendor specific software, which essentially is a black box[224], or in open-source software tools.[229] Different centroiding procedures have been discussed for instance by Urban et al.[227] and more recently in a review by Renner et al.[97] Typical steps in centroiding are outlined below for a recorded scan (mass spectrum in profile mode) taking the algorithm described by Samanipour et al. as an example:[225]

0. Three parameters need to be defined by the user: Nominal resolution, minimum intensity threshold, and goodness of fit threshold
1. Locating absolute maximum as starting point and selecting a tentative mass window of two times the nominal resolution
2. The algorithm finds inside the selected mass window the points to the left and to the right of the maximum, at which the intensity has dropped to half the intensity compared to the maximum value, to select a refined mass window
3. The mass window selected in 2. is smoothed using a moving average filter with a fixed three-point smoothing window
4. A three parameter Gaussian peak model is fitted to the smoothed mass window
5. Goodness-of-fit is calculated between the Gaussian peak model and the smoothed data. If the Goodness-of-fit is higher than the user-defined threshold the algorithm moves to Step 6, otherwise the current mass window is discarded
6. The mean of the Gaussian peak model is compared to the position of the maximum value found in Step 1. If the difference is smaller than half the user-defined nominal resolution, the algorithm moves on to Step 7, otherwise the current mass window is discarded
7. Centroid position is calculated as the average of the mean of the modeled Gaussian peak and the maximum position found in Step 1. The height of the centroid is scaled by the area of the peak, and the window is flagged as processed

8. The algorithm goes back to Step 1 looking for the next largest intensity in the scan. The algorithm terminates if no datapoint with intensity above the user-defined threshold is left.

The algorithm described above was applied to centroid 20 scans of the simulated mass spectral profile data shown in **Figure 12A**. The resulting centroided raw data is shown in **Figure 12B**. The user-defined parameters for the centroiding algorithm need to be adjusted based on the instrument characteristics. Urban et al. show, that different mass analyzers produce different mass profile data, and that the assumption of Gaussian peak shapes does not necessarily hold for instance for TOF instruments.[227] Additionally, the final positions of the centroids can vary slightly across scans and therefore need to be aligned, as shown in **Figure 12C**. This step is also referred to as EIC construction. Alignment in the mass spectral direction is performed using either equidistant or non-equidistant binning.[226] In equidistant binning, the m/z-axis is partitioned in bins centered at specific m/z values and all m/z values within the boundaries of a respective bin are aligned to the m/z value of the center point. For example, let the bin size be 0.5 centered around the m/z value 121. Then m/z values 121.1, 120.9, 121.4, 120.6, and 120. 7 would be aligned to the m/z value of 121. The given example describes a simple rounding to nominal mass, which is commonly performed to align low resolution mass spectrometry data. Two main drawbacks are associated with equidistant binning, which are mostly relevant to high resolution mass data. The first is the risk of peak splitting, which means that the m/z values belonging to the same fragment are aligned to different m/z values which leads to distorted EIC shapes. The second drawback is, that the data compression is related to a loss of mass spectral resolution, which is especially a problem for high resolution data. Both drawbacks can be circumvented to some degree by non-equidistant binning algorithms like the region-of-interest (ROI) algorithm proposed by Tautenhahn et al.[230] The ROI procedure does not only align m/z values but also filters out data points that are below a user-defined threshold and discards EICs that have less consecutive scans than the user-defined peak widths.[230] Different implementations of the ROI algorithm are now being used inside many data analysis methods and workflows.[25,33,231] Alternative approaches for EIC construction have been published, such as a method by Aberg et al., which uses a procedure adapted from radar observation based on Kalman tracking.[232] Another method, proposed by Zhu et al., is based on hierarchical density clustering and requires fewer user-defined parameters compared to the ROI approach.[233] The ROI algorithm is conceptually illustrated in **Figure 13A**, with **Figure 13B** providing a simplified example to visualize the effect of ROI binning on the centroided raw data. A constructed EIC is also referred to as "feature", which is according to Tautenhahn et al. defined

as "*a bounded, two-dimensional (m/z and retention time) LC/MS signal*"[230], however, the same definition applies to GC-MS. Both, centroiding and alignment can potentially affect the data structure and produce artifacts due to the user-parametrized filter operations.

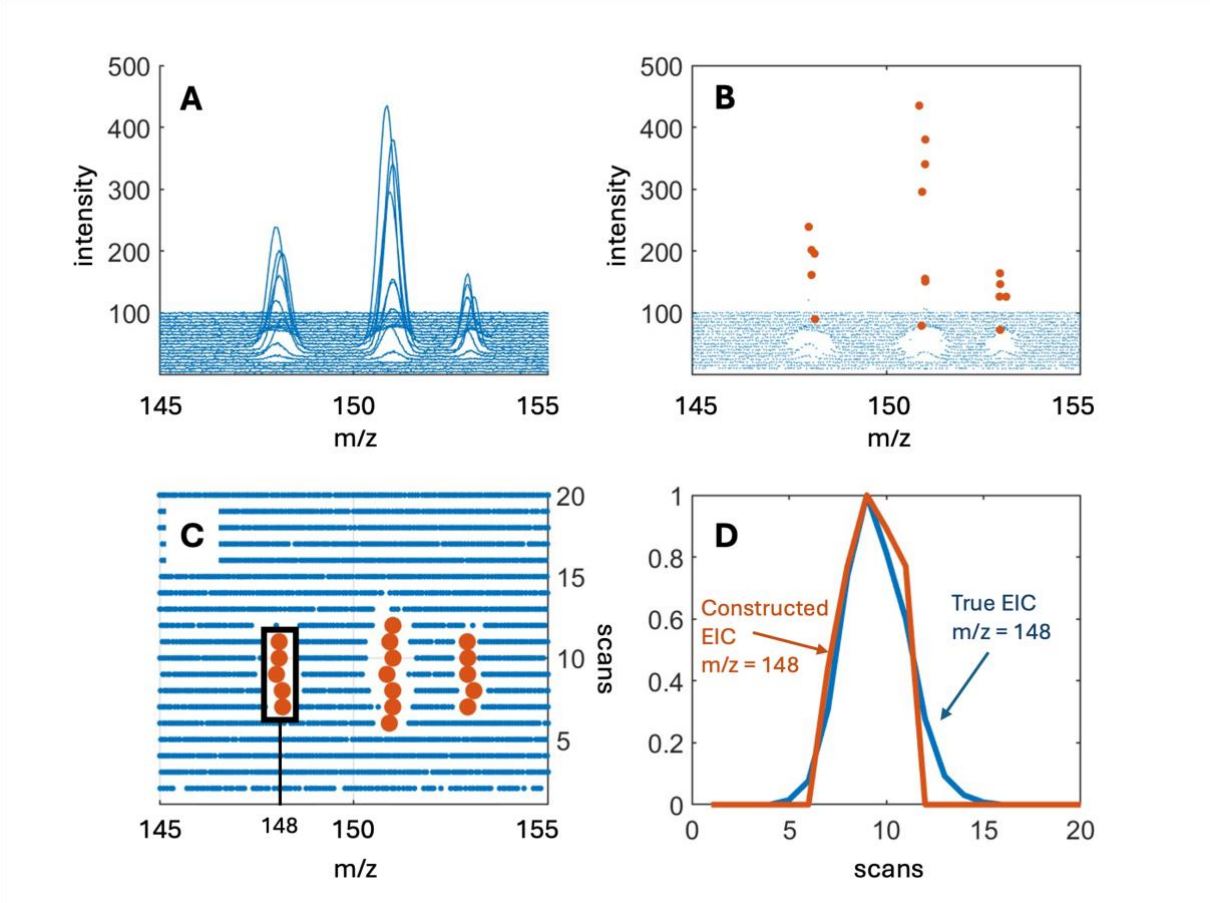


***Figure 12**: Visualization of the effect of centroiding. **A:** Discrete-continuous profile data of 20 simulated retention time scans with mass peaks at m/z 148, 151, and 153. **B:** Centroided data (using the algorithm described by Samanipour et al.[225]), with orange dots representing the positions of the centroids. **C:** Visualization of the fluctuation in m/z values of centroids. Mass alignment is required as indicated by the black box to assign the centroids to a common m/z value. **D:** Comparison of reconstructed EIC after mass alignment with the true simulated EIC. The border points of the reconstructed peak have been discarded because the intensities were below the minimum intensity threshold.*

For example, applying noise-thresholding during centroiding and ROI binning may remove low abundant scans from the edges of a peak (see **Figure 12D** or [Paper IV]), leading to distorted peak shapes. It may also occur that entire EICs will not pass the consecutive-scans-filter implemented in the ROI algorithm (see **Figure 13B**).[230] In [Paper IV], it has been demonstrated that smoothing can help to mitigate peak shape distortion.

Introducing artifacts at this early stage of data processing is particularly problematic because they can affect all subsequent data analysis steps, with a high risk of staying unnoticed.[227] Renner et al. discuss this problem in their critical review on data processing algorithms in non-

target screening (untargeted analysis).[97] They conclude that the large number of user-defined parameters in centroiding and EIC construction algorithms lead to a lack of comparability in untargeted studies.[97] Since important diagnostic information about the data quality, captured by the peak shape of mass spectra in profile mode, is lost after centroiding, Reuschenbach et al. developed a scoring method to evaluate the quality of centroids.[234] This should help assessing the uncertainty in the pre-processed data, and informs downstream statistical analysis.[234] Uncertainty estimation has also been investigated by Yu et al., Guo et al., and Zhang et al., who investigated the effect of "computational variation" caused by data pre-processing tools on the quantitative analytical results in LC-MS-based metabolomic studies.[235] Further, they developed machine learning and artificial intelligence (AI)-based methods to assess the quality of EICs and the uncertainty introduced by computational variation.[236,237]

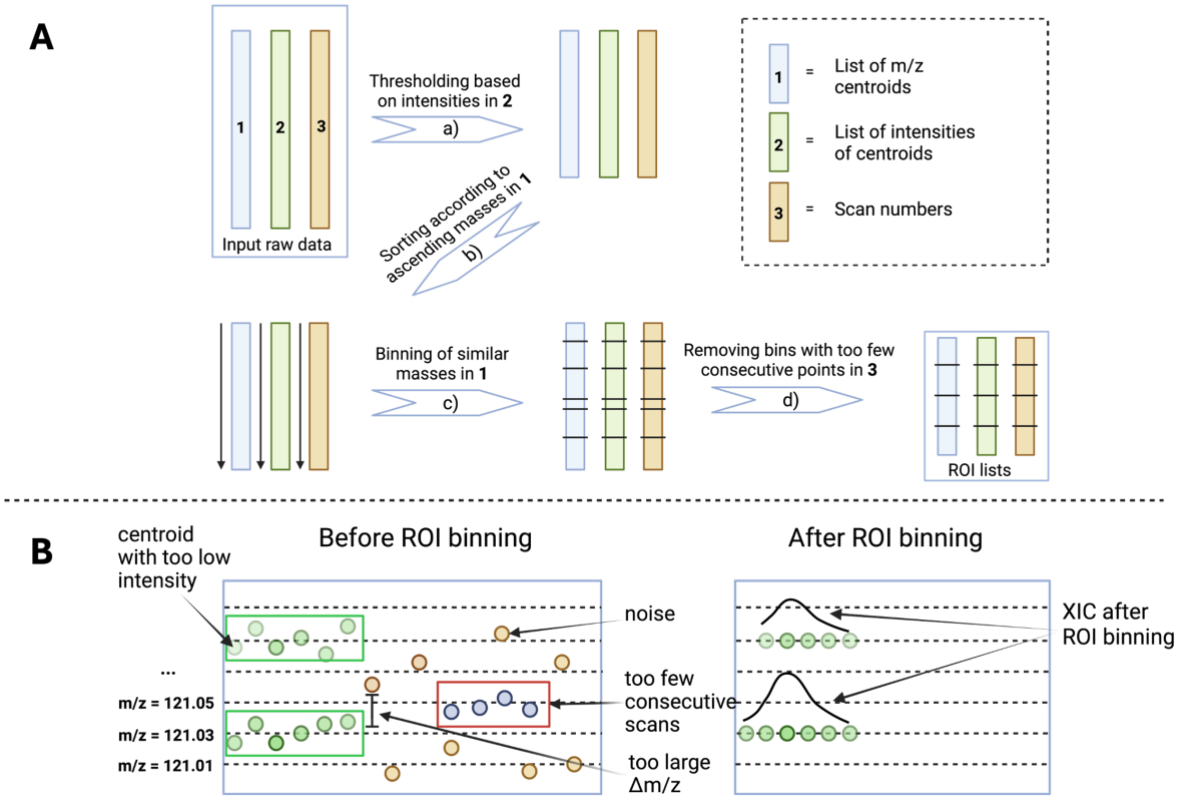


***Figure 13**: Visualization of the region-of-interest (ROI) algorithm. **A:** Schematic visualization of the algorithmic steps of the ROI algorithm. **B**: Multiple mass spectral scans before (left) and after applying ROI (right). Centroids in the boxes should be aligned, but some centroids are filtered out because they are below the noise threshold or because they consist of too few consecutive scans (red box). Noise that has passed the thresholding in the centroiding algorithm is filtered out. Created with BioRender.com*

4 Information extraction using curve-resolution and tensor decomposition methods

The symbiotic combination of data driven, and theory driven modelling is a major strength of the chemometric approach. Data driven modelling allows, in principle, for an unbiased exploration of the analytical data, while theory-driven modelling provides the foundation for causal interpretation of the results. Harald Martens called this the "abductive process", where inductive (data-driven) and deductive (theory-driven) modelling are combined to extract meaningful chemical information.[238] Additionally, physical and chemical knowledge can also guide the development of new chemometric methods based on hypotheses formulated for the structure of the measured data. This is important, if existing methods yield unsatisfactory results because they either make assumptions about the data structure that do not match the actual complexity, or they do not incorporate available a priori knowledge that could guide the model toward a more meaningful solution. Having a sufficient theoretical understanding of the nature of the data is perhaps the most important aspect in ensuring the application of the parsimonious principle in chemometric data analysis.[239]

In this Chapter, different chemometric models used for modelling chromatographic data are presented and explained. In Chapter 3 the foundational theory for understanding the structure of chromatographic data has been introduced. Based on this, model assumptions and limitations are discussed, considering the theoretical and practical aspects related to the chromatographic data.

Models such as parallel factor analysis (PARAFAC)[240], parallel factor analysis 2 (PARAFAC2)[241], and MCR-ALS[242] became popular in chemometrics for deconvoluting chromatographic datasets, largely due to the work of Tauler et al. (1994) and Bro et al. (1999), who demonstrated their applications in LC with diode array and excitation-emission detection.[243–245] However, even earlier, methods like generalized rank annihilation (GRAM), evolving factor analysis (EFA) and non-iterative MCR (self-modeling curve resolution) were used to deconvolute overlapped chromatographic signals.[246–248] Non-iterative MCR methods, such as SIMPLISMA[249], are now often used to generate starting values for iterative MCR-ALS, which has become more widely used due to the development of sophisticated constraints.[250]

Since a comprehensive review of all these methods is beyond the scope of this work, the focus will be on MCR-ALS, PARAFAC and PARAFAC2, as they are most commonly used methods today. The following Sections will describe and discuss the use of these methods for modelling

data from GC-MS, GC×GC-MS, LC-MS(/MS) and LC×LC-MS(/MS). Fundamental concepts introduced in the first Section on GC-MS will be revisited and expanded in subsequent Sections on GC×GC-MS, LC-MS(/MS), and LC×LC-MS(/MS).

4.1 Definition of "information extraction from chromatographic data"

The goal of information extraction is to convert chromatographic raw data into peak tables which contain the (relative) concentration and the chemical identity of all relevant analytes. In an unsupervised, untargeted study, this peak table should ideally comprehend all peaks, since it has not been decided *a priori*, which peaks are relevant. In a supervised untargeted study, it is sufficient to only extract information of peaks that explain differences between sample classes. In a semi-targeted study, the peak table should summarize the extracted information of the suspected compounds.

To practically explain how this can be achieved, it is helpful to first revisit the components that constitute a recorded chromatographic signal.[251] In **Figure 14A**, a small retention time window of the TIC signal from a GC-MS measurement is shown. The signal in **Figure 14A** is the sum of all the components shown in **Figures 14B-D**. Specifically, **Figures 14B1-B3** show the EICs of three different analytes eluting in the retention time window, representing the chemical information that should be extracted. **Figures 14C-D** show the chromatographic baseline signal (low-frequency noise) and high-frequency noise from the detector, which must be separated to extract "clean" chemical information. Various algorithms exist for removing the low-frequency and high-frequency noise components, which are referred to as baseline correction and smoothing methods. Some of these algorithms will be discussed in Chapter 5, but for more comprehensive reviews of baseline correction and smoothing methods, the reader is referred to the literature.[252,253] The defining characteristic of curve resolution and tensor decomposition methods is their ability to mathematically decompose different parts of the chromatographic signal into distinct components.[26] Thus, extraction of chemical information is achieved by selecting the "chemically meaningful" components (**Figure 14B1-B3**) and discarding the uninformative components (**Figure 14C**).

Baccolo et al. proposed a convolutional neural network-based peak classification method to distinguish peaks and select "chemically meaningful" components from PARAFAC2 decompositions.[254,255] While low-frequency noise contributions (e.g. baseline signal) are typically captured by additional components, high-frequency noise often remains undescribed by the model and can be found in the model residuals. Problems may arise if the contribution of high- and low-frequency noise to the recorded signal is much higher than the contribution of

the analyte signal. In such low SNR situations, pre-processing steps like smoothing and baseline correction may still need to be incorporated into the decomposition procedure [Paper III-IV]. Different decomposition methods have different requirements (e.g., bilinear, or multilinear)[242,256] regarding the structure of the modelled data. To extract chemically meaningful components, it is crucial that these requirements are met by the structure of the experimental data. A more detailed mathematical description of the decomposition process and the different requirements will be provided in the following Sections.

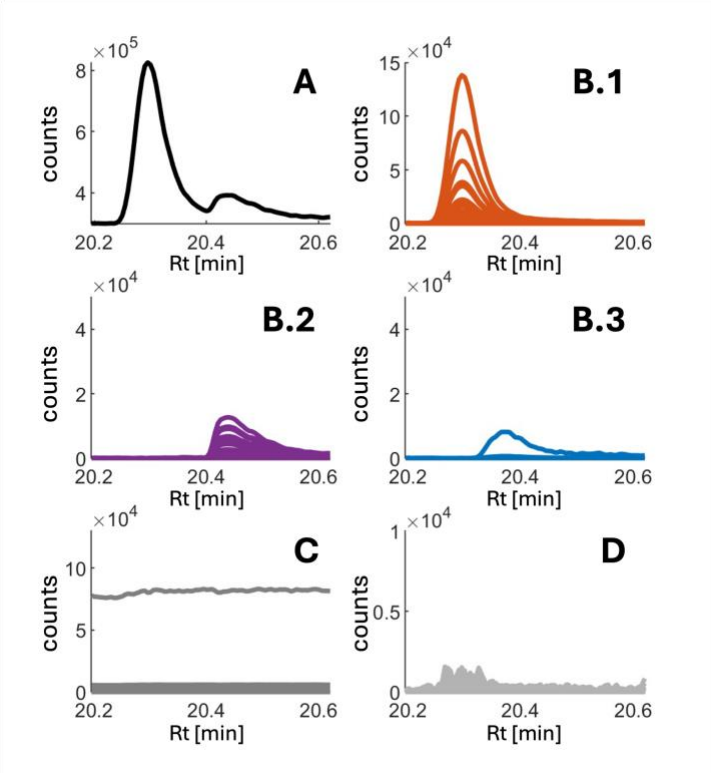


Figure 14*: Parts of the chromatographic signal.* ***A****: Retention time window of a total ion chromatogram (TIC),* ***B.1-3****: EICs of three different analytes that are co-eluting in the retention time window.* ***C****: Baseline (low-frequency noise).* ***D****: High-frequency detector noise.*

The extracted chemical information (**Figure 14B1-B3**) is presented in a slightly different form in **Figure 15**. Instead of displaying the individual EICs for each analyte, the sum of all EICs (analyte TIC) is shown for each analyte, along with its respective mass spectrum. One key piece of information is the area under the curve of the analyte TIC signal, which serves as a measure of the analyte's relative concentration. The peak areas of all analytes across samples can be used in downstream univariate or multivariate statistical analyses e.g., to test if the abundance of certain analytes is higher in one group of samples than in another. The other critical piece of information is the mass spectrum, which can be used to determine the chemical structure of the analyte (annotation), for example, by comparison to mass spectral databases.[170] In summary, information extraction refers to the process of isolating analyte specific signals from interfering

signals to obtain both, relative-quantitative information (peak area) and qualitative information (analyte mass spectrum).

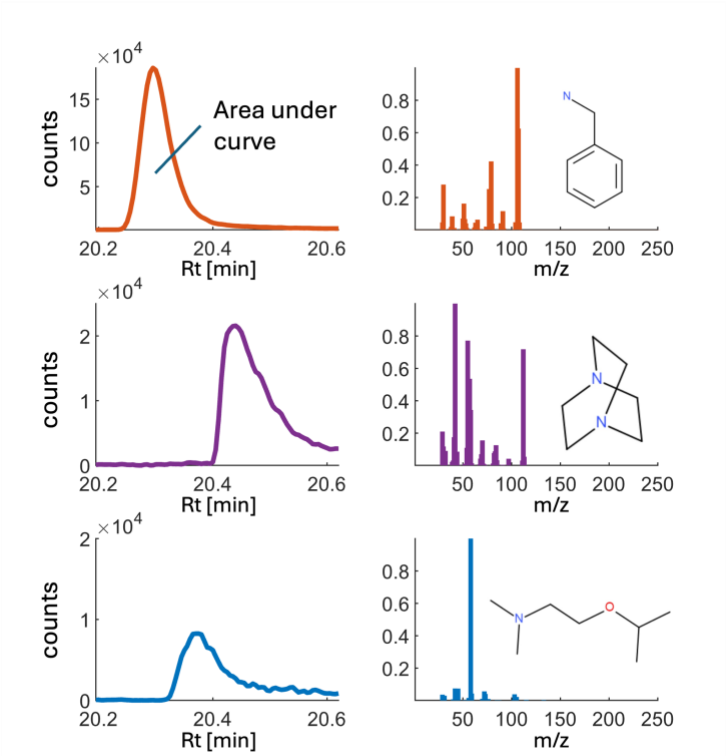


***Figure 15**: Different visualization of the chemical information shown in **Figure 14B.1-3**. **Left**: Total ion chromatograms (TIC) of the analyte signals, from which peak areas can be extracted as relative quantitative information. **Right**: Mass spectra of the respective analytes which can be used for compound identification.*

When it comes to the implementation of information extraction methods, different approaches need to be distinguished. Information extraction methods can either operate on small retention time windows of the dataset to extract chemical information incrementally[88,257–259], or in an "all-at-once" approach to extract chemical information from the entire dataset at once.[260] For curve resolution and tensor decomposition methods, the "all-at-once" approach is only feasible for less complex datasets and if there are not too large concentration differences between the analytes of interest. The "one-at-a-time" approach is on the other hand well suited to extract also information of low abundant peaks from complex datasets.[26] However, this approach requires the selection of multiple retention time windows and can be computationally demanding.[255] The retention time window that has been analyzed in **Figure 14** and **Figure 15** is shown in **Figure 16** in comparison to the TIC of the entire GC-MS measurement.

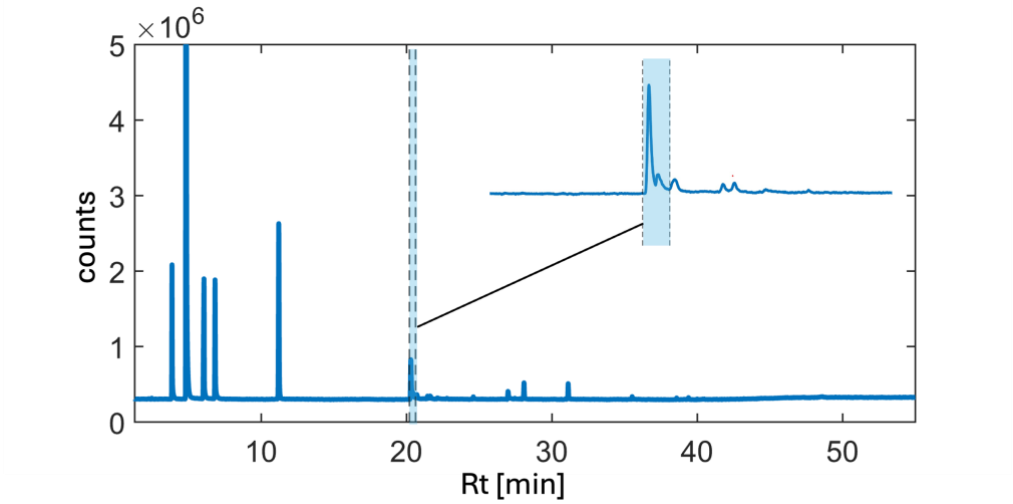


***Figure 16**: Visualization of the total ion chromatogram of an entire GC-MS measurement and the respective retention time window (light blue) that has been analyzed in **Figure 14** and **Figure 15**.*

Information extraction approaches can be categorized into those that leverage information across multiple measurements[25,88,257], which will in this work be referred to as "batch mode", and those that process measurements individually.[34,258,259,261] In this context, "batch mode" does not imply the programmatic implementation (e.g., using parallel computing), but refers to the algorithmic processing of the data (e.g., using efficient arrangements of the measured data). Sections 4.2.2 to 4.2.4 will demonstrate that curve resolution and tensor decomposition methods can natively leverage information across multiple samples and are therefore usually applied in batch mode, which offers several advantages discussed in Section 5.3.[262–265]

Another important aspect of implementing information extraction methods is the need to specify the number of components into which the dataset should be decomposed. This is especially crucial for curve resolution and tensor decomposition methods, but it also applies to some of the methods discussed in Chapter 5. For curve resolution and tensor decomposition methods rank analysis is typically used to estimate the number of components. Specifically, the number of significant singular values describing the structured variance of the chromatographic dataset is used to determine the "chemical rank" of a dataset.[266–268] Alternatively, the number of components can be determined post analysis using diagnostic tools that evaluate the properties of decompositions with different component numbers.[269–271] More recently, peak detection and deep learning methods have been proposed to determine the number of components.[258,272]

4.2 Information extraction from gas chromatography coupled to mass spectrometry

4.2.1 Multivariate curve resolution and rotational ambiguity

A single GC-MS measurement can be expressed in the form of a matrix $\mathbf{X}$ $(I \times J)$, with I denoting the number of scans at consecutive time points and J denoting the number of mass channels. Under the assumption that a GC-MS measurement has a bilinear data structure, $\mathbf{X}$ $(I \times J)$ can be decomposed into two factor matrices $\mathbf{C}$ $(I \times R)$ and $\mathbf{S}$ $(J \times R)$, where $\mathbf{C}$ is holding the elution profiles and $\mathbf{S}$ is holding the mass spectra of R resolved components. This is the MCR model, defined by Eq. 8.[242] The term "components" comprehends analytes, co-eluting species, or baseline contributions, as has been described previously. The part of the data $\mathbf{X}$ that is not described by $\mathbf{C}$ and $\mathbf{S}$ remains in the residual matrix $\mathbf{E}$ $(I \times J)$.

In this context it is important to emphasize that in MCR-ALS, it is generally assumed that each analyte has a rank-one contribution and can be modeled by one component $r \in \{1, \dots, R\}$, meaning it is represented by the outer product of one elution profile $\mathbf{c}_r$ with one mass spectrum $\mathbf{s}_r^{\mathrm{T}}$. This has been referred to as the principle of "chemical rank".[268] Consequentially, it is implicitly assumed that the analyte mass spectra recorded at different retention time scans are consistent and independent of the concentration changes described by the elution profiles. In Sections 3.4.1 and 3.4.2 situations have been described in which these assumptions are violated. For instance, spectral skewing can lead to a systematic change in the mass spectra, which is proportional to the concentration changes between different retention time scans. However, besides the work of Samokhin, the effect of spectral skewing has not yet been systematically investigated in the chemometric literature.[201] For the rest of this Section it is assumed that the mass spectra recorded at different retention time scans are independent of concentration changes, unless the opposite is stated explicitly.

Equation 8: $$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E}$$

The loss function of the ALS algorithm can be formulated according to Eq. 9., showing the case of MCR-ALS with non-negativity constraints on the elution profiles and the spectra. For a non-negative, bilinear dataset containing independently, identically distributed noise, MCR-ALS is optimal.[242]

Equation 9:
$$L(\mathbf{C},\mathbf{S}) = \left| |\mathbf{X} - \mathbf{C}\mathbf{S}^{\mathrm{T}}| \right|_F^2 \; s.t.$$
$$c_{nr} \geq 0 \; \forall \; n \in \{1, \dots, I\}, r \in \{1, \dots, R\}$$
$$s_{mr} \geq 0 \; \forall \; m \in \{1, \dots, J\}, r \in \{1, \dots, R\}$$

The use of constraints plays a pivotal role in MCR-ALS to incorporate physical and chemical *a priori* knowledge into the data modelling. For instance, it is physically meaningful, to assume non-negativity for mass spectra and elution profiles in GC-MS data, as the incident ion beam hitting the photomultiplier will not produce negative signals (see **Figure 8C**). Furthermore, the use of constraints is important to reduce rotational ambiguity of MCR solutions, which will be explained in the following.

The MCR model suffers from three ambiguities which are scale ambiguity, permutation ambiguity and rotational ambiguity.[242,273] Scale ambiguity and permutation ambiguity are also present in PARAFAC and PARAFAC2 models and generally do not have large practical relevance.[265] Scale ambiguity is usually handled by normalizing all factor matrices except one (in MCR either $\mathbf{C}$ or $\mathbf{S}$ is normalized), and thereby keeping the scale information in the other factor matrix.

The permutation ambiguity refers to the lack of inherent ranking or ordering of components in methods like MCR or PARAFAC/PARAFAC2. In PCA, the components are ordered based on their contribution to the total variance and thus have a clear ranking. However, in MCR or PARAFAC/PARAFAC2, the specific order of components can vary depending on the initial values and constraints applied to each factor. This ambiguity arises because there is no predefined rule or ranking that dictates which component should appear first, second, or third, unlike in PCA.

Unlike the other two ambiguities, rotational ambiguity occurs only in MCR and not in PARAFAC/PARAFAC2 and has large practical implications.[28,242,273] Mathematically, rotational ambiguity means that in general no unique solution, but a range of solutions minimizes $L(\mathbf{C},\mathbf{S})$. This can be shown by Eq. 10., in which $\mathbf{T}$ is a transformation matrix and $\mathbf{C}_{\mathrm{A}}$ and $\mathbf{S}_{\mathrm{A}}^{\mathrm{T}}$ are rotated versions of $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$, which provide the same fit. Practically, this means that there is no guarantee that a solution found for Eq. 9. describes meaningful elution profiles and mass spectra.

Equation 10:
$$\mathbf{X} \approx \mathbf{C}\mathbf{S}^{\mathrm{T}}$$
$$\mathbf{C}\mathbf{S}^{\mathrm{T}} = (\mathbf{C}\mathbf{T})(\mathbf{T}^{-1}\mathbf{S}^{\mathrm{T}}) = \mathbf{C}_{\mathrm{A}}\mathbf{S}_{\mathrm{A}}^{\mathrm{T}}$$

Manne formulated two theorems defining conditions under which unique profiles can be recovered from LC-DAD data with MCR.[268] The first theorem states that "*if all interfering compounds that appear inside the concentration window of a given analyte also appear outside this window, it is possible to calculate the concentration profile of the analyte", and the second theorem, "If for every interferent the concentration window of the analyte has a sub-window where the interferent is absent, then it is possible to calculate the spectrum of the analyte.".*[268] To be consistent in the terminology used throughout this work, the terms "*concentration window*" and "*concentration profile*" will be referred to as "elution window" and "elution profile" in the following. Under the assumption that the mass spectra recorded at different scans are not affected by time-dependent concentration changes, Manne's theorems can be applied to the GC-MS data structure.

Three different GC-MS datasets have been simulated to graphically illustrate situations where the theorems formulated by Manne may or may not be fulfilled and to examine the consequences for the MCR-ALS solutions. In the simulated data, elution profiles are modeled as Gaussian profiles and the mass spectra are represented as vectors of sparse, randomly generated positive numbers, scaled so that their maximum intensity is one. In all examples, the mass spectra of the analyte and the interferent share three fragments, to introduce some correlation.

For each of the simulated cases A, B and C, displayed in **Figure 17**, the simulated profiles and spectra are shown, along with the feasible solutions for profiles and spectra, that minimize $L(\mathbf{C}, \mathbf{S})$ under non-negativity constraints. The polygon inflation algorithm[274] implemented in the FACPACK software v.2.0 (https://www.math.uni-rostock.de/facpack/Downloads.html, 03.07.2024) was used for the calculation of the feasible solutions.

Figure 17A shows a simulated case in which both theorems are fulfilled. The elution profiles of an analyte (blue curve) and an interferent (orange curve) are displayed, with the elution window of the analyte indicated by vertical black dotted lines. The purple-colored region shows that the interferent also elutes outside the analyte's elution window, satisfying the first theorem. The second theorem is satisfied because there is a sub-window within the elution window where only the analyte elutes (green-colored region). As a result, only one feasible solution (abbreviated with FS in the header of **Figure 17**) exists for the elution profiles and the mass spectra. In **Figure 17B**, a case is shown where the interferent elutes entirely within the elution window of the analyte. In this situation, the first theorem is violated for the analyte and the second theorem is violated for the interferent. Consequentially, only the mass spectrum of the analyte and the shape of the interferent's elution profile are uniquely defined. Although the

shape of the elution profile of the interferent is uniquely defined, its scale is not, meaning that scaled versions of the uniquely defined elution profile exist within the range of feasible solutions. **Figure 17C** depicts a case recently studied by Carabajal et al.[275], where, instead of a co-eluting interferent, a linear background is shown. Since the background is also present outside the concentration window of the analyte, the first theorem holds, and the shape of the elution profile is uniquely defined (same situation as for the interferent in **Figure 17B**). On the other hand, because the background profile is present throughout the entire elution window of the analyte, the second theorem is violated for the analyte, resulting in rotational ambiguity in the mass spectral estimates of the analyte. With no region where only the analyte is present without the background, there is no unique solution for the background profile. Nevertheless, since there are regions where only the background is present, the background's mass spectrum is uniquely defined according to Manne's second theorem. In their study, Carabajal et al. demonstrated that reducing the rotational ambiguity introduced by a background signal is challenging, even when applying multiple active constraints, such as non-negativity, unimodality, area correlation and correspondence.[275] However, they proposed a background interpolation constraint that showed promising results.[275]

From this simple experiment, it can be concluded that rotational ambiguity complicates the extraction of reliable concentrations and mass spectra when dealing with a single GC-MS measurement. To decrease rotational ambiguity, the MCR-ALS routine typically incorporates additional constraints beyond non-negativity.

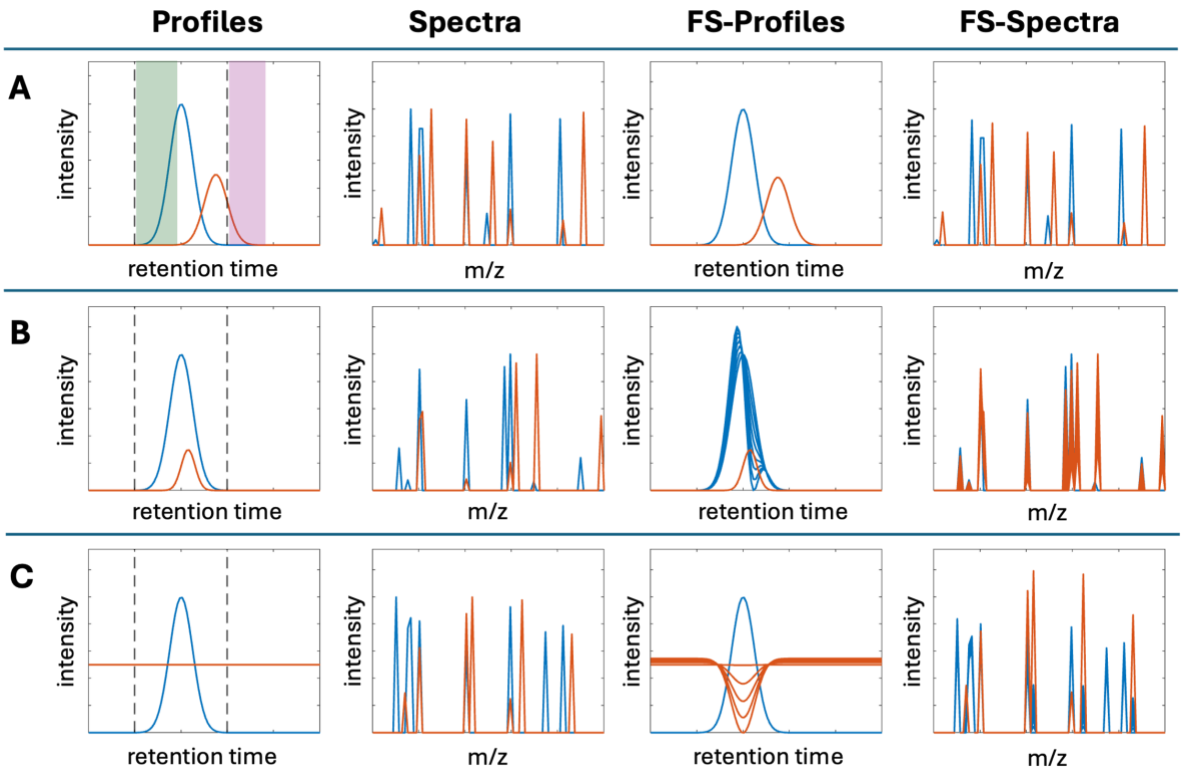


***Figure 17**: Visualization of different chromatographic scenarios and their implications for the multivariate curve resolution (MCR) model. In case **A**, both of Manne's resolution theorems are fulfilled and the MCR model has a unique solution (see FS- Profiles and FS-Spectra). In **B** and **C,** the resolution theorems are violated and the MCR model suffers from rotational ambiguity. Hence, multiple solutions exist which are visualized in the columns FS-Profiles and FS-Spectra. For more information, see description in the main text.*

4.2.2 Extended multivariate curve resolution and multiset analysis

In untargeted studies, it is more common to have multiple GC-MS measurements, with the challenge being the efficient analysis of large datasets.[257] Fortunately, the multiset approach in MCR provides an efficient framework for analyzing batches of GC-MS measurements.[243,244,262] Moreover, having more data can help to reduce rotational ambiguity in MCR solutions and more options for applying constraints exist.[263,266,276]

As described earlier, a single GC-MS measurement can be expressed as a matrix $\mathbf{X}$ $(I \times J)$, with I denoting the number of scans at consecutive time points and J denoting the number of mass channels. Therefore, a set of K measurements is a set of K matrices $\mathbf{X}_{(K)}$ with identical dimensions $(I \times J)$, under the assumption that all measurements have the same number of scans in retention time and cover the same m/z range. There are four different ways, how a set of matrices $\mathbf{X}_{(K)}$ can be arranged, depending on the objective of the modelling.

The first option is to concatenate the K matrices along the retention time mode to form an augmented matrix $\mathbf{X}_{I,aug}$ with dimensions $(IK \times J)$. The second option is to concatenate all

matrices $\mathbf{X}_{(K)}$ along the mass spectral mode to form the augmented matrix $\mathbf{X}_{J,aug}$ with dimensions $(I \times JK)$. The third approach is to stack the K column vectors of the vectorized matrices $\mathbf{X}_{(K)}$ to form the unfolded matrix $\mathbf{X}_{K,aug}$ with dimensions $(IJ \times K)$. Finally, the fourth option is to stack the matrices $\mathbf{X}_{(K)}$ to form a tensorial data structure $\underline{\mathbf{X}}^{(3)}$ with dimensions $(I \times J \times K)$. Modelling tensorial data structures requires the use of tensor decomposition methods such as PARAFAC and PARAFAC2, which will be revisited in the next Section. Figure 18 shows schematically how the different matrices $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$, and $\mathbf{X}_{(3)}$ are concatenated to form the augmented matrices $\mathbf{X}_{I,aug}$ and $\mathbf{X}_{J,aug}$.

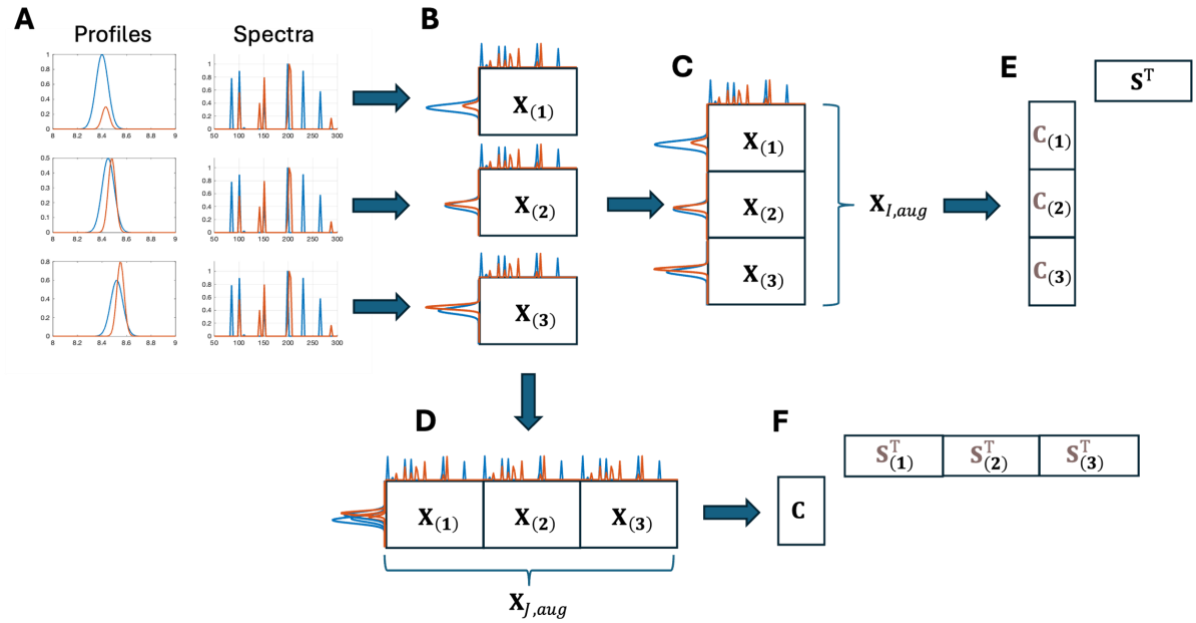


***Figure 18**: Examples of different multisets. **A**: Three different samples constructed from scaled versions of the elution profiles from **Figure 17B**, mass spectra are the same. **B:** Datasets $\boldsymbol{X}_{(k=1,2,3)}$ are constructed by the outer products of elution profiles with the respective mass spectra. **C:** Construction of the row-wise augmented matrix $\boldsymbol{X}_{I,aug}$. **D:** Construction of the column-wise augmented matrix $\boldsymbol{X}_{J,aug}$. **E:** Bilinear decomposition of $\boldsymbol{X}_{I,aug}$ provides estimates of the concatenated elution profiles $\boldsymbol{C}_{(k=1,2,3)}$ and the mass spectra. **F:** Bilinear decomposition of $\boldsymbol{X}_{J,aug}$ provides estimates of the mean elution profiles and the concatenated spectra. Since $\boldsymbol{X}_{(k=1,2,3)}$ are assumed to have a shared column space but not a shared row space, **B→C →E**. is the preferred augmentation / decomposition strategy as opposed to **B→D →F**.*

In multiset analysis of GC-MS data, the commonly used data structure is $\mathbf{X}_{I,aug}$, since the column space of the matrices $\mathbf{X}_{(K)}$ is assumed to be shared, while the row space is not.[262] Phenomenologically, this implies that an analyte present in $k \in \{1, \ldots, K\}$ measurements will produce the same mass spectrum in each of the k measurements, but the elution profiles may vary depending on the experimental conditions. Although the assumption of constant mass spectra can be debated for special cases, it is reasonable to assume that the variability in elution profiles is larger than the variability of the mass spectra. The variability in elution profiles is

due to retention time shift and other factors discussed in Section 3.2. For completeness, $\mathbf{X}_{J,aug}$ refers to the case where the row space of the matrices $\mathbf{X}_{(K)}$ is shared, but not the column space. This arrangement is typical when a set of samples is analyzed using different analytical techniques (e.g., NIR, Raman, NMR), and is also known as data fusion or multi-block analysis.[262,277,278] This type of augmentation will be revisited in Section 4.4 in the context of tandem mass spectrometry. Comparing the rotational ambiguity in single-measurement cases to multiple-measurement cases, it has been shown that multiple measurements alone can reduce rotational ambiguity.[263] Additionally, in the multiple-sample case, constraints that are not feasible in single-measurement scenarios can be applied. Notably, the trilinearity constraint can be used to completely eliminate rotational ambiguity, resulting in a unique solution. [244,262,276]

4.2.3 Multi-way analysis

The trilinearity constraint in MCR can be viewed as a special implementation of the PARAFAC model.[279] The PARAFAC model assumes a multilinear data structure of the form $\underline{\mathbf{X}}^{(p>2)}$. In chemometrics, the analysis of data structures in the form $\underline{\mathbf{X}}^{(p>2)}$ is referred to as multi-way analysis and comprehends a large family of methods.[280,281] As described before, a set of GC-MS measurements can be arranged to form a tensor $\underline{\mathbf{X}}^{(3)}$ with dimensions $(I \times J \times K)$, for which a PARAFAC decomposition exists (see Figure 19A-D).

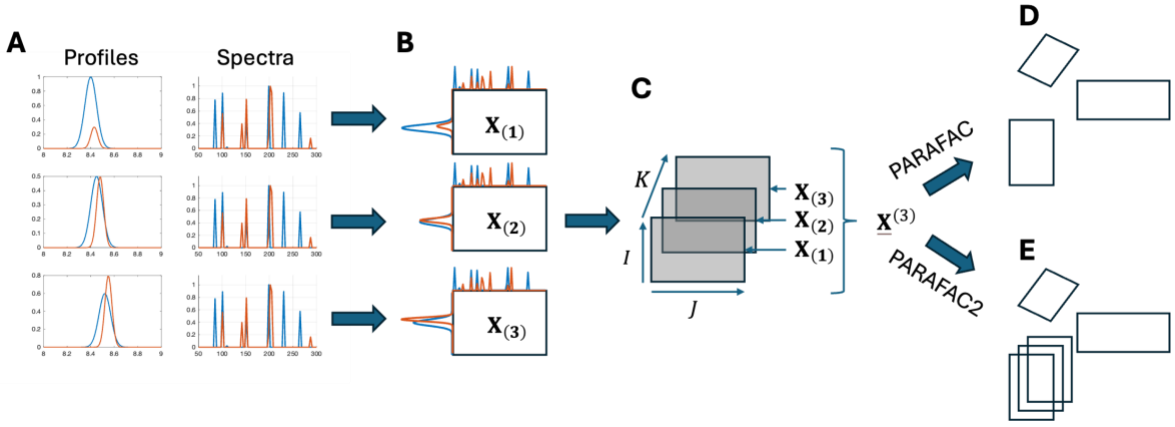


***Figure 19**: Multi-way data structure of multiple GC-MS measurements. **A** and **B** are the same as in Figure 18, showing the simulated data and the data matrices $\mathbf{X}_{(k=1,2,3)}$ of individual measurements, respectively. **C**: Stacking the individual matrices $\mathbf{X}_{(k=1,2,3)}$ provides the three-way array $\underline{\mathbf{X}}^{(3)}$. **D**: If the matrices $\mathbf{X}_{(k=1,2,3)}$ share the same row-space and column-space, a PARAFAC model can accurately describe the chemical information in $\underline{\mathbf{X}}^{(3)}$. **E**: If instead elution profiles are shifted or peak shape changes occur, the less constrained PARAFAC2 model will provide more meaningful factors.*

The motivation for using the PARAFAC model lies in its uniqueness under mild conditions - conditions that are often met in practical applications.[256,282,283] Consequently, the PARAFAC model, under these mild conditions, does not suffer from rotational ambiguity and provides

meaningful estimates of underlying elution profiles and mass spectra if the modeled data has a multilinear structure.[245,256,282,283] A multilinear data structure is present if each of the matrices $\mathbf{X}_{(k)}$, which have been stacked to form $\underline{\mathbf{X}}^{(3)}$, can be modeled according to Eq. 11. In this equation, $\mathbf{A}$ $(I \times R)$ is a factor matrix containing the estimates of the elution profiles, $\mathbf{B}$ $(J \times R)$ is a factor matrix holding the mass spectral estimates, and $\mathbf{D}_{(\mathrm{k})}$ is a $(R \times R)$ diagonal matrix, holding weights for the kth sample. The matrix $\mathbf{D}_{(\mathrm{k})}$ is containing the kth row of a factor matrix $\mathbf{D}$ in its diagonal elements (in literature often denoted as $\mathbf{C}$, but $\mathbf{D}$ is chosen to avoid confusion with the MCR notation).

Equation 11 $$\mathbf{X}_{(k)} = \mathbf{A}\mathbf{D}_{(\mathrm{k})}\mathbf{B}^{\mathrm{T}} + \mathbf{E}_{(\mathrm{k})}, \qquad k \in \{1, \dots, K\}$$

Different algorithms exist for calculating the PARAFAC model[281,284,285]; however, within the chemometrics community, the most widely used algorithm today is the ALS algorithm, originally proposed by Harshman and popularized by Bro.[240,265] The loss function for the PARAFAC-ALS model can be formulated according to Eq. 12:

Equation 12: $$L(\mathbf{A}, \mathbf{B}, \mathbf{D}) = \sum_{k=1}^{K} \left|\left|\mathbf{X}_{(k)} - \mathbf{A}\mathbf{D}_{(\mathrm{k})}\mathbf{B}^{\mathrm{T}}\right|\right|_F^2$$

Equations 11 and 12 make it clear that, for the PARAFAC model to hold, both the mass spectra and elution profiles must be consistent across the set of K samples.[264,265] Mathematically, this requires that all matrices $\mathbf{X}_{(k)}$ share the same row and column space. This is a strong assumption, given the previous discussion on artifacts in row-and column space (see discussion in Chapter 3.1 and 3.4). **Figure 18** illustrates a scenario in which all $\mathbf{X}_{(k)}$ share the same column space but not the same row space, due to peak shifts and shape changes. This is assumed to be an appropriate data arrangement because changes in elution profiles are typically larger than changes in mass spectra. Hence, a model that relaxes the strong assumptions of the PARAFAC model on the elution profiles while maintaining essential uniqueness would be an improvement over both the PARAFAC and the MCR model.

Bro et al. proposed using the less constrained PARAFAC2 model for modelling chromatographic data.[264,286] The PARAFAC2 model provides unique solutions under mild conditions [241] and offers a significant advantage over the PARAFAC model by allowing elution profiles to vary from sample to sample to a certain extent. The mathematical formulation of the PARAFAC2 model is given by Eq. 13, in which $\mathbf{A}_{(\mathrm{k})}$ represents the elution profile estimates for the kth sample, $\mathbf{H}$ $(R \times R)$ is the matrix of cross products of the elution profile estimates in

the kth sample, and $\mathbf{P}_{(\mathrm{k})}$ $(I \times R)$ is an orthonormal projection matrix. The factor matrices $\mathbf{B}$ and $\mathbf{D}_{(\mathrm{K})}$ are the same as in Eq. 11 and 12.

Equation 13	$\mathbf{X}_{(k)} = \mathbf{A}_{(\mathrm{k})}\mathbf{D}_{(\mathrm{k})}\mathbf{B}^{\mathrm{T}} + \mathbf{E}_{(\mathrm{k})}$	$k \in \{1, \dots, K\}$
with	$\mathbf{H} = \mathbf{A}_{(\mathrm{k})}^{\mathrm{T}}\mathbf{A}_{(\mathrm{k})}$	$k \in \{1, \dots, K\}$
and	$\mathbf{A}_{(\mathrm{k})} = \mathbf{P}_{(\mathrm{k})}\mathbf{H}$	$k \in \{1, \dots, K\}$

Equation 13 shows that the PARAFAC2 model still assumes some common structure between the elution profiles by requiring constant cross-products $\mathbf{A}_{(\mathrm{k})}^{\mathrm{T}}\mathbf{A}_{(\mathrm{k})}$ across samples. A practical implication of this cross-product constraint is that only retention time shifts that preserve the degree of overlap and the retention order of neighboring peaks can be modeled (see **Figure 5** and discussion in Section 3.2.1). The same restriction applies, to the degree to which shape changes in elution profiles can be modeled.

Algorithms for calculating the PARAFAC2 model can be categorized into direct fitting[264] and indirect fitting[241,287] algorithms. Direct fitting algorithms fit the original data $\mathbf{X}_{(k)}$, while indirect fitting algorithms fit the cross-products matrix of the original data $\mathbf{X}_{(k)}^{\mathrm{T}}\mathbf{X}_{(k)}$. The direct fitting algorithm is preferred in chemometrics because it is computationally more efficient and allows constraints to be imposed on the factor matrices $\mathbf{D}$ and $\mathbf{B}$.[264] The direct fitting algorithm divides the minimization problem stated in Eq. 14 into a two-step procedure.

Equation 14:	$L(\mathbf{P}_{(\mathrm{k})}, \mathbf{H}, \mathbf{B}, \mathbf{D}) = \sum_{k=1}^{K} \left\| \mathbf{X}_{(k)} - \mathbf{P}_{(\mathrm{k})}\mathbf{H}\mathbf{D}_{(\mathrm{k})}\mathbf{B}^{\mathrm{T}} \right\|_{\mathrm{F}}^{2}$

In the first optimization step, $\mathbf{P}_{(\mathrm{k})}$ is found by the singular value decomposition of $\mathbf{H}\mathbf{D}_{(\mathrm{k})}\mathbf{B}^{\mathrm{T}}\mathbf{X}_{(k)}^{\mathrm{T}}$ for all $k \in \{1, \dots, K\}$. When all $\mathbf{P}_{(\mathrm{k})}$ have been estimated this way, $\mathbf{X}_{(k)}$ can be projected on $\mathbf{P}_{(\mathrm{k})}^{\mathrm{T}}$ to simplify the optimization problem stated in Eq. 14 to Eq. 15. The optimization problem formulated in Eq. 15 has now the same structure as Eq. 12, allowing the second optimization step to be performed by finding the PARAFAC solution for Eq. 15.

Equation 15:	$L(\mathbf{H}, \mathbf{B}, \mathbf{D}) = \sum_{k=1}^{K} \left\| \mathbf{P}_{(\mathrm{k})}^{\mathrm{T}}\mathbf{X}_{(k)} - \mathbf{H}\mathbf{D}_{(\mathrm{k})}\mathbf{B}^{\mathrm{T}} \right\|_{\mathrm{F}}^{2}$

From Eq. 13-15 it follows that the direct fitting algorithm does not explicitly optimize over the elution profiles $\mathbf{A}_{(\mathrm{k})}$. Instead, they are obtained from the hard coupling term $\mathbf{A}_{(\mathrm{k})} = \mathbf{P}_{(\mathrm{k})}\mathbf{H}$.

This has the implication, that constraints on the elution profiles (e.g., non-negativity or unimodality) need to be applied to the matrix product $\mathbf{P}_{(k)}\mathbf{H}$. It has been shown that this is not straightforward[288] and accordingly, no practical solution of constraining the elution profiles in the PARAFAC2 direct fitting framework exists.

Generally, because of the essential uniqueness of PARAFAC and PARAFAC2, constraints do not play such a paramount role in multi-way analysis as they do in bilinear modelling. Nevertheless, constraints are still useful, for improving convergence speed and stability, as well as for mitigating issues related to local minima solutions.[31] One prominent phenomenon related to convergence instability is two-factor degeneracy.[289,290] The occurrence of two-factor degeneracy is the consequence of the fact that, in general, no solution of a rank-*R* PARAFAC model exists.[291] Conditions under which two-factor degeneracies are likely to occur, have been empirically investigated[292,293] and formally analyzed.[289,290] Furthermore, Lim provided proof for the empirical observation that a globally unique solution exists, if a non-negative PARAFAC model is fitted to a non-negative tensor.[279], meaning two-factor degeneracies do not occur in this scenario.[279,291] Moreover, Yu and Bro demonstrated that applying constraints and using sophisticated initialization schemes can largely reduce the occurrence of local minima solutions in PARAFAC2 models.[31] Hence, a model allowing for constraints in all modes is highly favorable from a practical perspective.

The PARAFAC2 flexible coupling algorithm proposed by Cohen and Bro, allows for the implementation of constraints on the elution profiles.[288] The innovation of this algorithm is that it indirectly imposes the cross-products constraint (which is important to maintain uniqueness) by means of penalized regression, thereby optimizing directly over $\mathbf{A}_{(k)}$ in the ALS routine.[288] The loss function of the flexible coupling algorithm is formulated according to Eq. 16.

Equation 16: $$L\left(\mathbf{A}_{(k)}, \mathbf{P}_{(k)}, \mathbf{H}, \mathbf{B}, \mathbf{D}\right) = \sum_{k=1}^{K} \left|\left|\mathbf{X}_{(k)} - \mathbf{A}_{(k)}\mathbf{D}_{(k)}\mathbf{B}^{\mathrm{T}}\right|\right|_{\mathrm{F}}^{2} + \mu_{(k)} \left|\left|\mathbf{A}_{(k)} - \mathbf{P}_{(k)}\mathbf{H}\right|\right|_{\mathrm{F}}^{2}$$

A drawback of the flexible coupling algorithm is that it introduces an additional parameter $\mu_{(k)}$ which needs to be tuned. Additionally, because $\mathbf{X}_{(k)}$ is not projected on $\mathbf{P}_{(k)}$ for the calculation of the least squares solution of $\mathbf{A}_{(k)}$, the computational requirements are higher for the flexible coupling algorithm compared to the direct fitting algorithm. To address the slow convergence speed of both the direct fitting and the flexible coupling algorithm, Yu et al. proposed the PARASIAS model. This model attempts to convert the data structure of $\mathbf{X}_{(k)}$ into a PARAFAC-like problem using a time-frequency domain transfer. This approach would be significantly

faster to compute than a PARAFAC2 model and would allow for constraints in all modes.[294] The reasoning behind the domain transfer is that by fitting the model to its frequency domain representation, shifts in elution profiles could be effectively removed due to the shift-invariant property of the amplitude spectra. If shape changes in the elution profiles across samples contribute minimally to the row rank of $\mathbf{X}_{(k)}$, the frequency domain representation of all $\mathbf{X}_{(k)}$ should share the same row space and column space. However, there is a conceptual problem with the PARASIAS model. To explain this problem, a time domain signal $f(t)$ is considered which is the sum of the time domain signals of two co-eluting analytes $f_1(t)$ and $f_1(t)$. The Fourier transform of $f(t)$ is denoted as $\hat{f}(\omega)$, which is given by the sum of the Fourier transforms of $f_1(t)$ and $f_2(t)$, due to the linearity of the Fourier transform (Eq. 17).:

Equation 17: $$\hat{f}(\omega) = \hat{f}_1(\omega) + \hat{f}_2(\omega)$$

The power spectrum $P(\omega)$ is calculated according to Eq. 18 as the product of the $\hat{f}(\omega)$ with its complex conjugate $\bar{\hat{f}}(\omega)$.

Equation 18: $$P(\omega) = \hat{f}(\omega)\bar{\hat{f}}(\omega)$$

Substituting in Eq. 18 $\hat{f}(\omega)$ for $\left(\hat{f}_1(\omega) + \hat{f}_2(\omega)\right)$ and $\bar{\hat{f}}(\omega)$ for $\left(\bar{\hat{f}}_1(\omega) + \bar{\hat{f}}_2(\omega)\right)$, leads to the expression in Eq. 19, which shows that the calculation of the power spectrum generates the cross terms $\hat{f}_1(\omega)\bar{\hat{f}}_2(\omega)$ and $\hat{f}_2(\omega)\bar{\hat{f}}_1(\omega)$.

Equation 19: $$\hat{f}(\omega)\bar{\hat{f}}(\omega) = \hat{f}_1(\omega)\bar{\hat{f}}_1(\omega)+\hat{f}_1(\omega)\bar{\hat{f}}_2(\omega)+\hat{f}_2(\omega)\bar{\hat{f}}_1(\omega)+\hat{f}_2(\omega)\bar{\hat{f}}_2(\omega)$$

Hence, if the cross-terms have non-zero contribution in Eq. 19, the Fourier transformed raw data $\mathbf{X}_{(k)}$ will suffer from rank inflation. This observation has also been acknowledged by Yu et al. who state that eventually more factors are required for the PARASIAS model to achieve factor resolutions comparable to the PARAFAC2 model.[294] Since the primary purpose of deconvolution algorithms like PARASIAS is to resolve co-eluting peaks, the rank inflation caused by co-eluting signals makes it less practical for GC-MS applications.

4.2.4 Extended multivariate curve resolution and shift-invariant trilinearity

Lim showed that non-negative PARAFAC and multilinear NMF (MCR) are equivalent.[279] Since non-negative versions of PARAFAC and PARAFAC2 are most meaningful for GC-MS data, the multilinear MCR framework[295] provides a viable alternative modelling approach. The implementation of trilinearity constrained MCR-ALS is schematically shown in **Figure 20A-C**. However, trilinearity constrained MCR makes the same assumptions as the PARAFAC model, limiting its application in the presence of chromatographic artifacts. To address this, a flexible implementation of the trilinearity constraint has been proposed by Zhang et al., for modelling chromatographic data with shifted elution profiles.[266] The method synchronizes the estimated elution profiles $\mathbf{C}$ prior to enforcing trilinearity.[266] Thereby, accurate estimates for elution profiles $\mathbf{C}$ and mass spectra $\mathbf{S}^{\mathrm{T}}$ as well as unique solutions are achieved even if elution profiles are shifted across different $\mathbf{X}_{(k)}$.[266]

In [Paper I], the idea of synchronizing the estimated elution profiles $\mathbf{C}$ instead of the raw data $\mathbf{X}_{(k)}$ has been combined with the idea of using the time-frequency domain transfer for establishing shift-invariance[294], to generalize the flexible trilinearity constraint[266] to a shift-invariant trilinearity (SIT) constraint. The conceptual idea behind the flexible trilinearity implementation and the SIT implementation is depicted in **Figure 20D**. Furthermore, it was demonstrated in [Paper I] that the SIT constraint provides significantly faster convergence than direct fitting and flexible coupling versions of PARAFAC2 (20 to 30 times faster). Additionally, SIT showed better factor resolution in situations of low peak intensity and high background signal, which has been reported to be challenge for PARAFAC2[198] and extended MCR[275] alike. The SIT model offers a promising alternative to PARAFAC2 because it can accommodate arbitrary shifts, including those that alter the degree of overlap or the retention order of neighboring peaks. One limitation of the SIT model is, however, that it cannot model shape changes of elution profiles occurring across samples. In [Paper II], a case is presented in which the peak of 1-(2-propenyloxy)-2-propanol shows concentration dependent shape changes across samples.

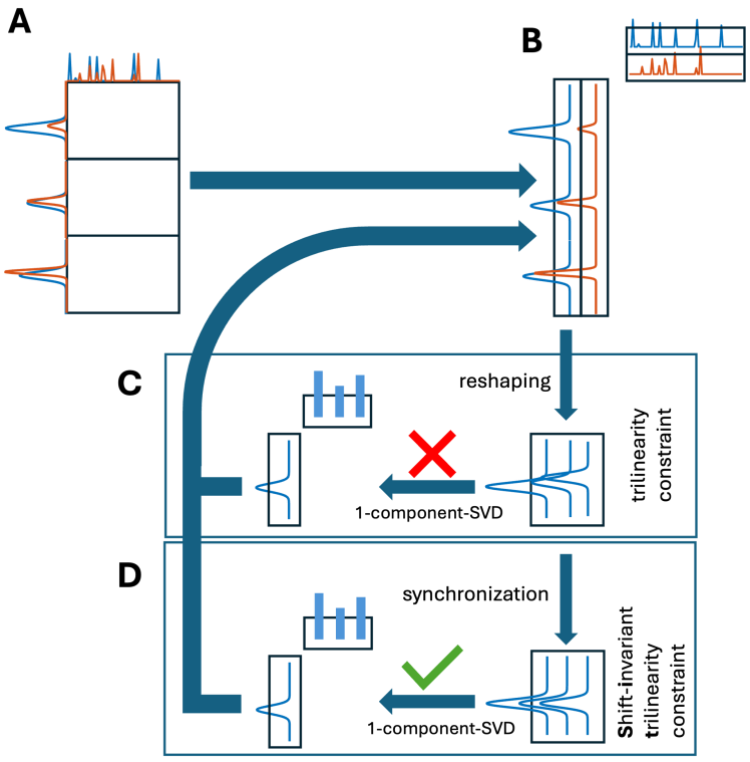


***Figure 20**: Visualization of trilinearity and shift-invariant trilinearity constrained MCR. **A:** Set of concatenated GC-MS measurements ($\boldsymbol{X}_{I,aug}$) is decomposed into a bilinear set of concatenated elution profiles and analyte specific mass spectra shown in (**B**). **C:** Each vector of concatenated elution profiles (only shown for blue to aid visualization) is constrained to impose trilinearity by reconstructing the matrix of concatenated elution profiles with a one-component SVD model. This reconstruction will not be accurate, if the individual elution profiles are shifted. **D**: In shift-invariant trilinearity, elution profiles are synchronized prior to reconstructing them with the one-component SVD model.*

The general root-causes for peak skewing were discussed in Section 3.2.3. In GC-MS peak shapes of alcohols, phenols, organic acids, and amines are particularly sensitive to column properties and conditions (film thickness, activity) as well as analyte concentration.[296–298] The SIT model exhibited larger errors in modelling the peak shapes of 1-(2-propenyloxy)-2-propanol in the dataset presented in [Paper II]. Conversely, flexible coupling PARAFAC2 and non-negativity constrained MCR provided accurate models. A relaxed version of SIT called shift-invariant soft trilinearity (SIST) was proposed which provides more flexibility in modelling peak shapes that are not only shifted but are also subject to shape changes. The flexibility of the SIST method is controlled by the number of principal components that are used for the reconstruction of the elution profiles. Thus, the SIST method, equivalently to the flexible coupling PARAFAC2 introduces a tunable hyperparameter. Despite the additional flexibility of the SIST method, it performs comparably to the more rigid SIT method, if the GC-MS data has a shifted tri-linear data structure. This was exemplified with a second dataset in [Paper II], which resembled a low abundant peak and high background signal. On the second

dataset, SIT and SIST provided better factor resolution than flexible coupling PARAFAC2 and non-negativity constrained MCR, which suffered from local minima and rotational ambiguity, respectively.

4.3 Information extraction from comprehensive two-dimensional gas chromatography coupled to mass spectrometry

The data structure of GC×GC-MS measurements is more complex than the data structure of GC-MS and different scenarios need to be considered for modelling this type of data. A single GC×GC-MS measurement has the structure of a higher-order tensor $\underline{\mathbf{X}}^{(3)}$, with dimensions $(I \times J \times K)$, and I denoting modulations in the first retention dimension (^{1}D), J denoting scans in the second retention dimension (^{2}D), and K denoting mass scans (m/z). Alternatively, the tensor $\underline{\mathbf{X}}^{(3)}$ can be unfolded along ^{1}D to give the augmented matrix $\mathbf{X}_{I,aug}$ with dimensions $(IJ \times K)$. The unfolding of $\underline{\mathbf{X}}^{(3)}$ is exemplified in **Figure 21A** for the kth slice denoted with $\mathbf{X}_{(k)}$ (two-dimensional EIC), which gives $\mathbf{x}_{I,aug}$ $(Ij \times 1)$ as the kth column of $\mathbf{X}_{I,aug}$.

Under ideal conditions, the PARAFAC model is well-suited for analyzing a single GC×GC-MS measurement, as has been shown by Hoggard et al. for both targeted and untargeted analysis.[299–301] Ideal conditions in this context mean that a given analyte eluting over several consecutive modulations in ^{1}D has the same retention time in each of the modulations in the second retention dimension ^{2}D. In this scenario, depicted in **Figure 21B**, the two-dimensional EICs of the analyte form two-dimensional Gaussians with zero covariance. If the different EICs have this form, they can be modeled by the outer product of two vectors which are scaled according to the abundance of a given mass fragment.

Deviations from the ideal conditions are caused by retention time shifts along ^{2}D, which result from temperature programming in ^{1}D.[23,302,303] Although the temperature in the second retention dimension ^{2}D can be considered constant during a given modulation, it increases in successive modulations to match the temperature ramp in the first retention dimension ^{1}D.[302] The consequence this temperature increase in successive modulations is that an analyte eluting over multiple modulations will be less retained, causing its retention time to decrease.[23,302,303] As shown in **Figure 21C**, under these conditions, the two-dimensional EICs take the form of Gaussians with non-zero covariance, which cannot be accurately reconstructed by the outer product of two vectors.

Therefore, the PARAFAC model will yield less accurate representations of the true elution profiles and mass spectra.[302] Pinkerton et al. investigated quantitatively the relationship between

retention time shift in ^{2}D ($\Delta_2^R t$), peak width in ^{2}D ($_b^2 W$), and the error of the PARAFAC model.[302] However, the PARAFAC2 model is better suited for modelling GC×GC-MS data if $\Delta_2^R t$ is not neglectable, as it can handle certain shifts. Using PARAFAC2 for modelling GC×GC-MS data has been proposed e.g., by Skov et al.[303]

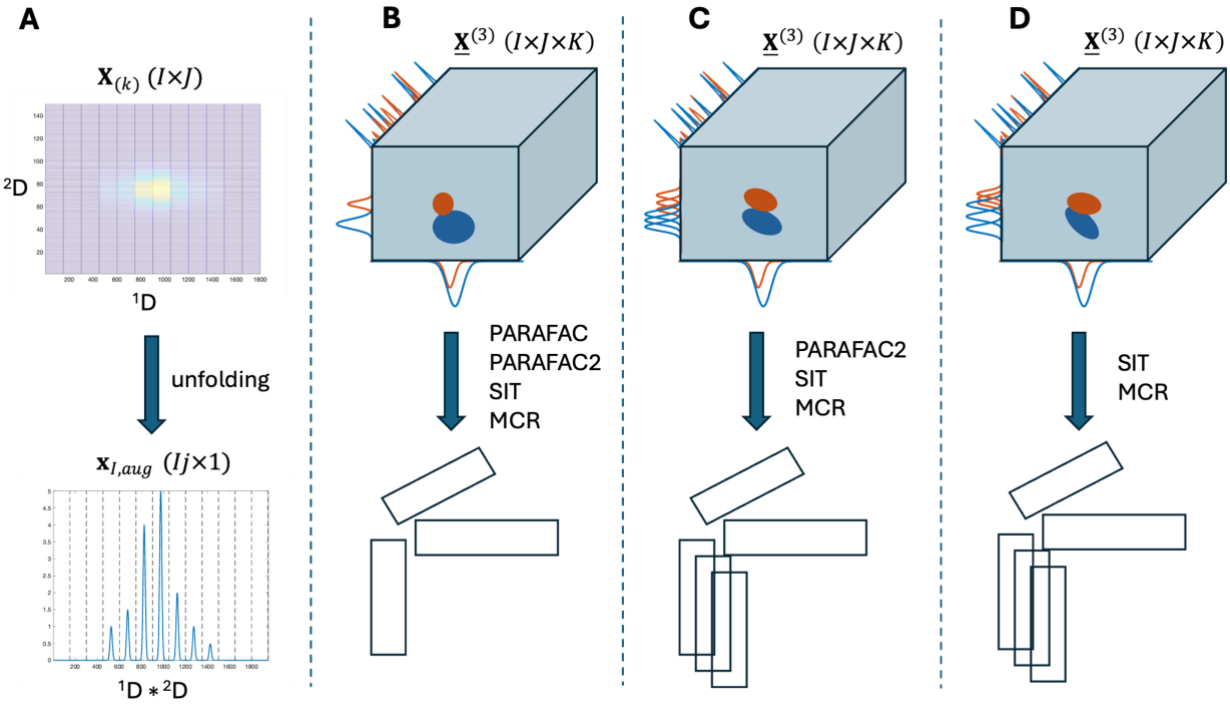


***Figure 21**: Visualization of different GC×GC-MS data structures. **A**: Unfolding of a 2D-EIC, **B**: GC×GC-MS measurement with neglectable second-dimension retention time shift ($\Delta_2^R t_i$), **C**: GC×GC-MS measurements with considerable $\Delta_2^R t_i$, where both analytes are equally affected ($\Delta_2^R t_1 = \Delta_2^R t_2$), **D**: GC×GC-MS measurements with considerable, analyte specific $\Delta_2^R t_i$ where both analytes are unequally affected ($\Delta_2^R t_1 \neq \Delta_2^R t_2$).*

In the case depicted in **Figure 21C**, the retention of the two co-eluting analytes is similarly affected by an increased temperature ($\Delta_2^R t$ is the same for both), reflecting the scenario of linear retention time shifts, discussed in Section 3.2.1. However, if one analyte's retention behavior is more sensitive to temperature increases than the other's, non-linear retention time shifts would occur (as also discussed in Section 3.2.1). Although this scenario has not been specifically investigated by Pinkerton et al.[302] or Skov et al.[303], the equations presented in their works for estimating $\Delta_2^R t$ are useful for illustrating this example. In Eq. 20, k_m is the retention factor (Eq. 2 and 4), $_2^R t_m$ is the retention time of the mth analyte in ^{2}D, $_2^R t_0$ is the dead time of ^{2}D, ΔH_m is the specific vaporization enthalpy of the mth analyte, R is the Avogadro constant, T is the absolute temperature, and C_m is a term describing the contribution of analyte-column-interaction to the retention of the mth analyte, which is equivalent to $\left(\frac{\Delta S_m}{R}\right) + \ln(\beta)$ in Eq. 4.

Equation 20: $$\ln(k_m) = \ln\left(\frac{{}^R_2t_m - {}^R_2t_0}{{}^R_2t_0}\right) = \frac{\Delta H_m}{RT} + C_m$$

Pinkerton et al. derived from Eq. 20 an expression for the retention time shift $\Delta {}^R_2t_m$, occurring between the Nth and the $(N+1)$th modulation, which is given by Eq. 21:

Equation 21: $$\Delta {}^R_2t_m = \frac{-\Delta H_m}{R}\left(\frac{\Delta T}{T_{N+1} \times T_N}\right)(k_m \times {}^R_2t_0)$$

In Eq. 21 it is assumed that $\Delta H_m = f(T, p)$ is constant between two consecutive modulations, because the flow conditions are not changing drastically and ΔT can considered to be small.[302,303] Nevertheless, ΔH_m is an analyte specific quantity which can vary considerably, depending on the functional groups of an analyte.[304] Further, $\Delta {}^R_2t_m$ is not only linearly dependent on ΔH_m but also exponentially due to $k_m = e^{\left(\frac{\Delta H_m}{RT} + C_m\right)}$, according to Eq. 20. The described situation is illustrated in **Figure 21D**, in which $\Delta {}^R_2t_{blue}$ is larger than $\Delta {}^R_2t_{orange}$. One challenge with this type of data is that the cross-product constraint of the PARAFAC2 model is violated due to changing degrees of overlap between neighboring peaks in consecutive modulations. The extent to which this situation can occur and may impact the results of the PARAFAC2 model should be investigated in future research. As an alternative to the PARAFAC2 model, the SIT model may be preferable, as it makes no assumptions about the nature of shifts and provides unique solutions. Additionally, the extended version of bilinear MCR-ALS has also been proposed for modelling GC×GC-MS data.[23,305] Especially, the less rigid assumptions regarding the data structure make it a versatile tool for all the discussed cases (**Figure 21B-D**). However, rotational ambiguity may pose a problem as has been discussed in the Section on GC-MS.

In larger studies, multiple GC×GC-MS measurements may be produced, and analyzing these datasets jointly can be more efficient and informative, especially in identifying common and distinct analytes across groups of samples, such as in metabolomics studies. A set of multiple GC×GC-MS measurements can be represented as a four-way array $\underline{\mathbf{X}}^{(4)}$ with dimensions $(I \times J \times L \times K)$, or unfolded as the augmented matrix $\mathbf{X}_{I,aug}$ with dimensions $(IJL \times K)$, with L representing the samples and I, J, and K representing ^{1}D, ^{2}D, and mass scans as in the single measurement case. From a modelling perspective, the joint analysis of multiple samples adds complexity because retention time shifts in ^{1}D and ^{2}D occurring between samples need to be

considered. For clarification, the retention time shift occurring between samples will be referred to as inter-sample shift, while the previously discussed deterministic retention time shift in ^{2}D will be referred to as intra-sample shift. Although multiple concatenated GC×GC-MS measurements still follow a bilinear data structure, most shift-invariant or relaxed multilinear methods cannot natively model four-way data with shifts in two modes and require prior shift correction in at least one mode.[23,161,305] This is the case for the PARAFAC, PARAFAC2, SIT, and SIST models. Armstrong et al. published an extended version of the flexible coupling PARAFAC2 algorithm called PARAFAC2 x N, which can model shifts in more than one mode. This advancement overcomes the limitations of shift-invariant or relaxed multilinear methods previously mentioned.[216] The authors showed on simulated and real GC×GC-MS datasets that their PARAFAC2 x N can effectively handle intra- and inter-sample shifts.[216] As an alternative method, the shift-invariant multilinearity (SIML) model was recently proposed in [Paper III]. The SIML model is an extension of the SIT model[306] that is amendable for modelling e.g., four-way data with shifts in two modes.

The results of a comparison of SIML with PARAFAC2 x N and extended MCR-ALS for the deconvolution of simulated and real GC×GC-MS datasets showed that SIML is more robust than PARAFAC2 x N, especially at increased noise levels. Further, the uniqueness of the SIML model is an advantage compared to extended MCR-ALS, which suffered from rotational ambiguity in some of the data examples presented in [Paper III]. The combination of SIML with integrated wavelet denoising, implemented as an optional feature, allowed for the extraction of accurate mass spectra and elution profiles even at very low SNR.

4.4 Information extraction from liquid chromatography coupled to mass spectrometry and tandem mass spectrometry data

Depending on the analytical scope, different instrumental configurations and acquisition modes need to be considered when hyphenating liquid chromatography to mass spectrometric detectors, as outlined in Chapter 3. In a first approximation, data structures acquired from LC-MS measurements are comparable to those from GC-MS measurements. One LC-MS measurement can be represented as a matrix $\mathbf{X}$ $(I \times J)$, if the data has been measured at more than one mass channel, as is the case, for example, in scan mode. A dataset consisting of multiple LC-MS measurements $\mathbf{X}_{(k)}$ can be represented in the form of an augmented matrix $\mathbf{X}_{I,aug}$ with dimensions $(IK \times J)$ or in a tensor structure $\underline{\mathbf{X}}^{(3)}$, as is shown in **Figure 18** and

Figure 19. Therefore, the same considerations regarding the applicability of the different curve resolution and tensor decomposition generally apply.

However, since the retention mechanism in LC is more complex than in GC, peak positions and shapes are less reproducible in LC, across larger sets of samples.[307,308] Bortolato et al. and Arancibia et al. have shown for HPLC-DAD data that artifacts, such as non-linear retention time shift and peak broadening are leading to the violation of the cross-product constraint of the PARAFAC2 model and ultimately result in inaccurate quantitative results.[29,309] Given these findings, extended MCR-ALS or the SIT and SIST model could provide a better fit to the data structure. Nevertheless, Khakimov et al. successfully used the PARAFAC2 model in a plant metabolomics study for the deconvolution of complicated chromatographic signals in 17 selected subregions of the full dataset.[310] They reported that the extracted mass spectra showed high similarity to recorded mass spectra and that the additional information provided by the PARAFAC2 deconvolution aided the separation of target and control group in the downstream data analysis.[310] Further, the PARAFAC model has been applied to a set of 34 LC-MS measurements of herbal extracts for classification and extraction of diagnostic ions.[311] By applying the model to the full-scale data (as opposed to the more common windowed approach) the authors could identify the most characteristic triterpene glycoside groups by their characteristic ions.[311] However, comprehensive raw data pre-processing was required to compensate for deviations from trilinearity caused by retention time shifts.[311]

The use of the extended MCR-ALS for modelling LC-MS data was described by Farres et al. and by Navarro-Reig et al., evaluating the metabolic profile of *Saccharomyces cerevisiae* and the changes in rice metabolome induced by Cd and Cu exposure.[312,313] Farres and co-workers reported that rotational ambiguity was only a minor issue in their results due to the high selectivity of LC-MS.[312] However, they pointed out that co-eluting analytes with common molecular or product ions were subject to rotational ambiguity, consistent with Manne's resolution theorems.[268,312] In more recent studies a combination of the ROI procedure[230] (see Section 3.4.4 for more details on ROI) with subsequent MCR-ALS decomposition, denoted as ROIMCR, has been proposed and validated.[25,226,314] Using ROI instead of equidistant binning has the advantage that high resolution information can be stored in in a compact form. It was shown by Dalmau et al.[314] that the ROI matrix $\mathbf{X}$ $(I \times J)$ created from a measurement of lipid standards could reconstruct the original data with as few as 346 m/z values (number of columns of $\mathbf{X}$). Equidistant binning of the same dataset would have created 35.000 m/z values for a bin size 0.05 Da.[314] In previous studies, the ROI procedure has not been used but instead m/z values were rounded to nominal mass to achieve manageable data sizes.[310,312,313] Up to date, no study

has been published utilizing the benefits of ROI pre-processing in combination with the PARAFAC2, SIT or SIST method for LC-MS data analysis.

Today, the predominantly used instrumental setup for untargeted analysis is LC-MS/MS. The significant advantage of tandem mass spectrometry is that it provides simultaneous information about the exact molecular mass and characteristic product ions of an analyte. In LC-MS/MS measurements, different scenarios with respect to DDA and DIA need to be distinguished (compare Chapter 3, **Figure 9**). The data obtained from DDA measurements does not suffer largely from co-elution in the chromatographic domain as only selected molecular ions are fragmented and specific mass transitions monitored. Hence, the application of chemometric deconvolution is not as crucial. Conversely, deconvolution methods play a bigger role for the data analysis of DIA experiments. In DIA experiments, two mass scans are recorded at every retention time point. One scan is recorded at a low collision energy (MS^1) and one scan at a high collision energy (MS^2). Hence, DIA experiments are considered less biased and more comprehensive than DDA experiments which comes at the cost of a more complex data. Each LC-MS/MS measurement in DIA produces two matrices $\mathbf{X}_{(k,MS1)}$ and $\mathbf{X}_{(k,MS2)}$ of size $(I \times J_{MS1})$, and $(I \times J_{MS2})$, respectively, representing the MS^1 and MS^2 mass spectra measured at different retention times. Sets of multiple measurements can be augmented row-wise to create the augmented matrices $\mathbf{X}_{I,aug,MS1}$ with dimensions $(IK \times J_{MS1})$, and $\mathbf{X}_{I,aug,MS2}$ with dimensions $(IK \times J_{MS2})$. Alternatively, the individual matrices can be reshaped into the two tensors $\underline{\mathbf{X}}_{MS1}^{(3)}$ with dimensions $(I \times K \times J_{MS1})$, and $\underline{\mathbf{X}}_{MS2}^{(3)}$ with dimensions $(I \times K \times J_{MS2})$. It is important to mention, that due to the ROI procedure, all matrices $\mathbf{X}_{(k)}$ will have a different number of columns, i.e., $J_{MS1} \neq J_{MS2}$ and $J_{k=1,MS1} \neq J_{k=2,MS1}$. Hence, before augmenting two matrices $\mathbf{X}_{(1,MS1)}$ and $\mathbf{X}_{(2,MS1)}$ row-wise, it is important to create a common m/z-axis by merging the m/z values found within $\mathbf{X}_{(1,MS1)}$ and $\mathbf{X}_{(2,MS1)}$. This is schematically shown in **Figure 22A** for $\mathbf{X}_{(1,MS1)}$, $\mathbf{X}_{(2,MS1)}$, and $\mathbf{X}_{(3,MS1)}$ but the concept is the same for augmenting $\mathbf{X}_{(1,MS2)}$, $\mathbf{X}_{(2,MS2)}$, and $\mathbf{X}_{(3,MS2)}$.

Different strategies have been reported for how to model the MS^1 and MS^2 datasets jointly. In the work of Perez-Lopez, $\mathbf{X}_{I,aug,MS1}$ and $\mathbf{X}_{I,aug,MS2}$ were column-wise concatenated (**Figure 22B**) to provide the new super-augmented matrix $\mathbf{X}_{IJ,aug,MS1,MS2}$ of size $(IK \times [J_{MS1} + J_{MS2}])$, which was then decomposed using MCR-ALS.[260] In the work by Kronik et al., PARAFAC2 flexible coupling was used to model three different sets of environmental samples.[198] The models were fitted to data structures of type $\underline{\mathbf{X}}_{sum,MS1,MS2}^{(3)}$, created by summing $\underline{\mathbf{X}}_{MS1}^{(3)}$ and $\underline{\mathbf{X}}_{MS2}^{(3)}$, after equidistant binning of the mass axis to nominal resolution (**Figure 22C**).[198]

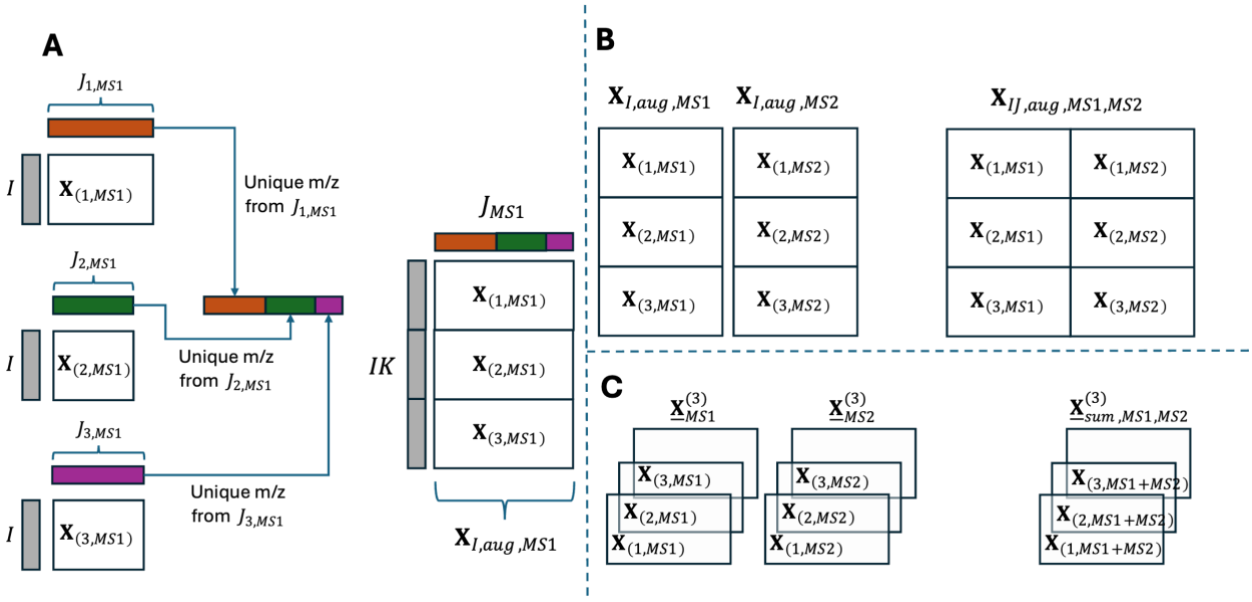


Figure 22**: Data handling strategies for LC-MS.* ***A: Construction of augmented matrices based on results from ROI binning. **B:** Creation of super-augmented matrix $\boldsymbol{X}_{IJ,aug,MS1,MS2}$ according to *Perez-Lopez et al.*[260]. ***C:*** *Construction of the summed tensor* $\underline{\boldsymbol{X}}^{(3)}_{sum,MS1,MS2}$ according to Kronik et al.[198]

Conversely to the findings cited previously[29,309], Kronik and co-workers reported only minor deviations from trilinearity in the retention dimension, which could be effectively handled by the flexible coupling PARAFAC2 model.[198] However, deviations from trilinearity have been observed in the mass spectral dimension.[198] Specifically, it has been observed that the mass spectra of three standard compounds were not consistent across the analyzed samples. The reason for this inconsistency could be appointed to the formation of adducts.[198] This is an important finding, because it does not only describe a violation of the trilinearity assumption but also of the bilinearity assumption.

For a matrix $\mathbf{X}_{I,aug}$, bilinearity can be assumed if all concatenated matrices $\mathbf{X}_{(k)}$ share the same column space, as has been described in Section 4.2.1. This implies that the mass spectrum of a given analyte needs to be consistent across samples. If instead the mass spectrum changes in some samples due to the formation of adduct fragments, the assumption of a shared column space is violated. In practice, more than one component will be required to model the part of the analyte mass spectrum that is common across samples and the part that changes across samples. A simple example has been simulated to illustrate this. **Figure 23A** shows concatenated EICs of three samples, in which one contains an EIC that is not present in the other two samples (orange colored). In **Figure 23B**, the mass spectrum of the first two samples (blue) is shown in comparison to the mass spectrum of the last sample (orange). The dataset shown in **Figure 23A** has been decomposed using a one-component and a two-component MCR-ALS model (non-negativity constraints only). The elution profile obtained from the one-

component model (blue) is displayed in **Figure 23C** overlayed with the true TIC profile (black dashed) of the data. While the modeled elution profile perfectly resembles the true TIC of the first two samples, a difference can be seen between the modeled elution profile and the true TIC for the third sample. In **Figure 23D**, the sum of the two elution profiles obtained from the two-component model is perfectly matching the overlayed true TIC profile. The example illustrates that adducts affect quantitative results, if it is not considered that analyte signals may need to be modeled by more than one component.

It has been pointed out in Section 3.4.3. that matrix effects and more specifically ion suppression and adduct formation are common problems in LC-ESI-MS(/MS). However, the influence of matrix effects in general has not been considered in the described applications of ROIMCR or tensor decomposition methods to LC-MS and LC-MS/MS.[25,226,260,310,312–314] besides the work of Kronik et al.[198]

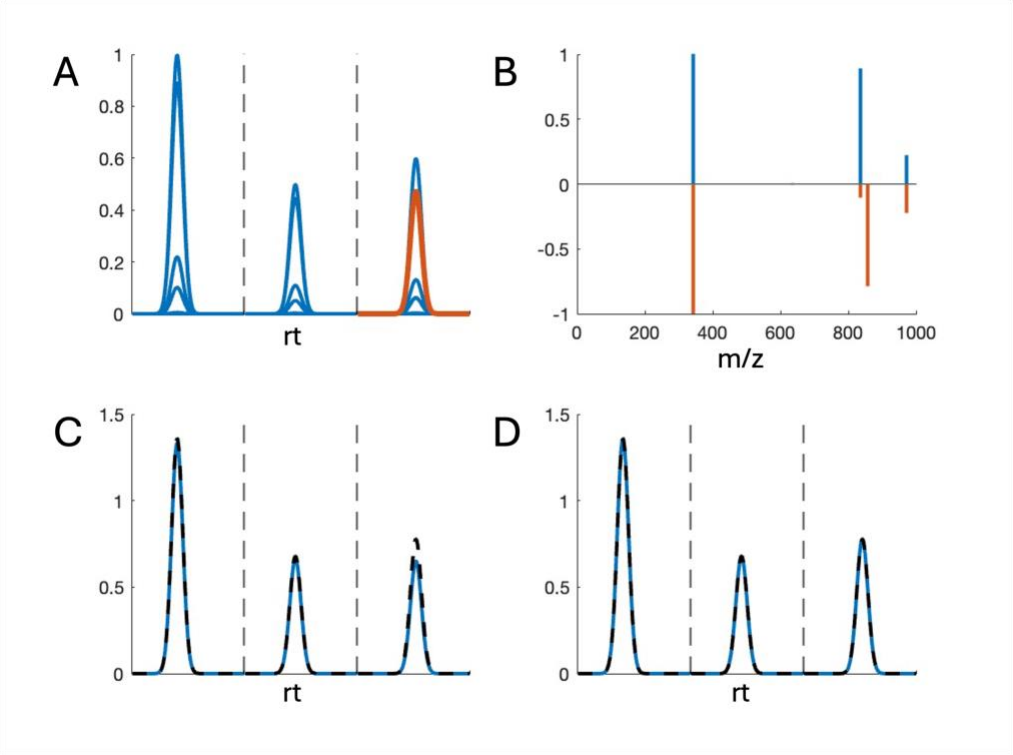


Figure 23*: Simulated example visualizing the effect of adduct formation in LC-MS. Retention times are omitted to aid visualization.* ***A:*** *Concatenated EICs of three samples. The third sample contains one EIC that is not present in the other two samples (orange-colored line).* ***B:*** *Mirror plot comparing the joint mass spectra of the first two samples (blue) with the mass spectrum of the third sample (orange).* ***C:*** *Elution profiles estimated with a one-component MCR model (blue) overlaid with true TIC profile (black dashed). Differences are present between modeled TIC and true TIC for the third sample.* ***D:*** *Sum of elution profiles estimated with two-component MCR model (blue) is perfectly matching the overlayed, true TIC profile (black dashed).*

4.5 Information extraction from comprehensive two-dimensional liquid chromatography coupled to mass spectrometry

The structure of comprehensive LC×LC-MS data is comparable to the structure of GC×GC-MS data in that single measurements present higher-order tensors $\underline{\mathbf{X}}^{(3)}$, with dimensions $(I \times J \times K)$. The tensor structure can be unfolded along the retention time mode to construct the augmented matrix $\mathbf{X}_{I,aug}$. $(IJ \times K)$. The ROI procedure can be used to compress the mass spectral data before constructing the augmented matrix.[315] Equivalently to GC×GC-MS data, multiple LC×LC-MS samples can be concatenated to produce a four-way array $\underline{\mathbf{X}}^{(4)}$ $(I \times J \times L \times K)$ or the respective super-augmented matrix $\mathbf{X}_{I,aug}$. $(IJL \times K)$. If tandem mass spectrometry is used, extended data structures incorporating both MS^1 and MS^2 data may be convenient to use, analogous to the one-dimensional LC-MS/MS case, describe in the previous Section.

Several scientific articles have explored the application of curve resolution and tensor decomposition methods to LC×LC-MS.[162,163,260,315–317] However, LC×LC is still a relatively young technology with research efforts focusing more on instrumental improvements and method development.[46,154,318] Despite this, immature data analysis workflows have been considered one of the main roadblocks hindering broader adoption of the technology.[46,162] Early applications of PARAFAC and MCR-ALS to LC×LC-UV data have been described by Fraga et al. and by Bailey et al.[319,320] Similarly to the GC×GC, it has been described that the second dimension retention time shift causes deviations from multilinearity for LC×LC data.[320] Furthermore, Navarro-Reig and co-workers investigated the structure of comprehensive LC×LC-MS data, concluding that retention time shifts between consecutive modulations as well as peak shape changes are causing deviations from trilinearity in $\underline{\mathbf{X}}^{(3)}$ $(I \times J \times K)$.[316] Their results suggest that MCR-ALS may provide the better structural model as compared to PARAFAC or PARAFAC2.[316] In a more recent study, Perez-Cova et al. supported the findings by Navarro-Reig et al., showing that sets of LC×LC-DAD-MS measurements do not adhere to the multilinearity assumption.[317] The SIST model could be a valuable alternative to the MCR model if rotational ambiguity limits the accurate resolution of mass spectra and elution profiles. This approach has been shown to effectively handle deviations from trilinearity caused by retention time shifts and peak shape changes in the chromatographic domain [Paper II].

Additionally, Perez-Cova and co-workers describe the quantification of different amino acids in commercial drug mixtures using MCR-ALS combined with non-negativity and area correlation constraints.[315,321] Despite the promising results reported by Perez-Cova et al., MCR-

ALS-based quantification strategies may be compromised by deviations from bilinearity caused by matrix effects, as discussed in Sections 4.4 and 3.4.3.[315] The proposed area constraint cannot handle the problem of mass spectra inconsistency across samples but only reduce the rotational ambiguity.[321]

One problem associated with the LC×LC separation is reduced sensitivity due to analyte dilution in the modulator.[46] Hence, trace compounds may not be detected or only detected at low SNR. Advanced modulation techniques such as stationary phase assisted modulation may help to overcome this issue in the future.[322] However, Ochoa et al. and Cain et al. showed that for GC×GC-MS data, curve resolution and tensor decomposition methods used for extracting clean mass spectra are challenged in low SNR situations.[30,72] The same problem has been recognized in suspect screening of different pollutants in a wastewater sample analyzed with LC×LC-MS/MS in DIA. In [Paper IV], an algorithmic workflow has been suggested for the extraction of clean MS^2 mass spectra from complex LC×LC-MS/MS data in suspect screening. After data compression with ROI, the proposed workflow uses a cosine similarity-based mass filtering to extract clean mass spectra. Co-elution in the data could be resolved more easily using MCR-ALS after mass filtering was applied to a pre-process the data as described in [Paper IV].

5 Alternative methods for information extraction

The chemometric methods for information extraction presented in Sections 4.2 to 4.4 have focused primarily on curve resolution and tensor decomposition. While some implementation aspects were mentioned in Section 4.1, many details were not addressed. In this Chapter, we discuss workflows and methods implemented in the most popular software tools for untargeted analysis, aiming to round out the discussion on information extraction methods. Specifically, this Chapter will: 1) introduce data analysis workflows commonly used for untargeted analysis at both conceptual and algorithmic levels, and 2) discuss differences between the workflows and methods presented in Chapters 4 and 5.

Section 5.1 introduces feature-based workflows widely used for untargeted GC-MS and LC-MS(/MS) data analysis, implemented in popular open-source software tools. Section 5.2 presents pixel-based and tile-based approaches used for the analysis of GC×GC-MS data. Finally, Section 5.3 compares the methods discussed in Chapters 4 and 5.

5.1 Feature-based data analysis workflows for one-dimensional chromatography coupled to mass spectrometry

To recall, a "feature" according to Tautenhahn et al. is defined as "*a bounded, two-dimensional (m/z and retention time) LC/MS signal*"[230], which is also applicable to GC-MS data. Feature-based workflows have been implemented in commercial or open-source software tools, packages or online-platforms that offer comprehensive data analysis capabilities, going beyond mere information extraction.[35,37,84–87] However, the focus in this Section will be solely on the information extraction aspect. Additionally, this review will be restricted to investigate workflows implemented in three of the most popular tools (based on the number of citations) for GC-MS and LC-MS(/MS) data analysis: MZmine 3[35], XCMS[85], and MS-DIAL 4[37]. For better readability, MZmine 3 and MS-DIAL 4 will be referred to as MZmine and MS-DIAL, with version numbers mentioned only when necessary for distinction. It is acknowledged that this will not provide a comprehensive overview over the state of the art, but the goal of this Section is to provide methodological examples of feature-based workflows. Feature-based workflows follow a conceptually similar procedure but differ in the specific algorithms used.[34,43,259,323] For instance, equidistant binning is used in MS-DIAL, (described in the

Supporting Information of Tsugawa et al.[34]) whereas different versions of ROI binning are employed in MZmine and XCMS.[229,323]

Figure 24 shows the workflows for MZmine, MS-DIAL, and XCMS with respect to the analysis of GC-MS and LC-MS(/MS) data. MZmine offers additional capabilities for modelling LC-IM-MS or MALDI data, which are not considered in the workflow shown below, as they are beyond the scope of this work.[35] Differences exist between the various toolboxes regarding the data structures they can handle. MS-DIAL is the only toolbox offering workflows for LC-MS/MS DIA data. MZmine does not offer deconvolution for LC-MS(/MS) data, whereas both MS-DIAL and XCMS do. The individual processing steps of feature-based workflows are presented below, and examples of implemented algorithms will be discussed in more detail. Broader reviews of functions and algorithms included in different feature-based workflows can be found in the literature.[87,324]

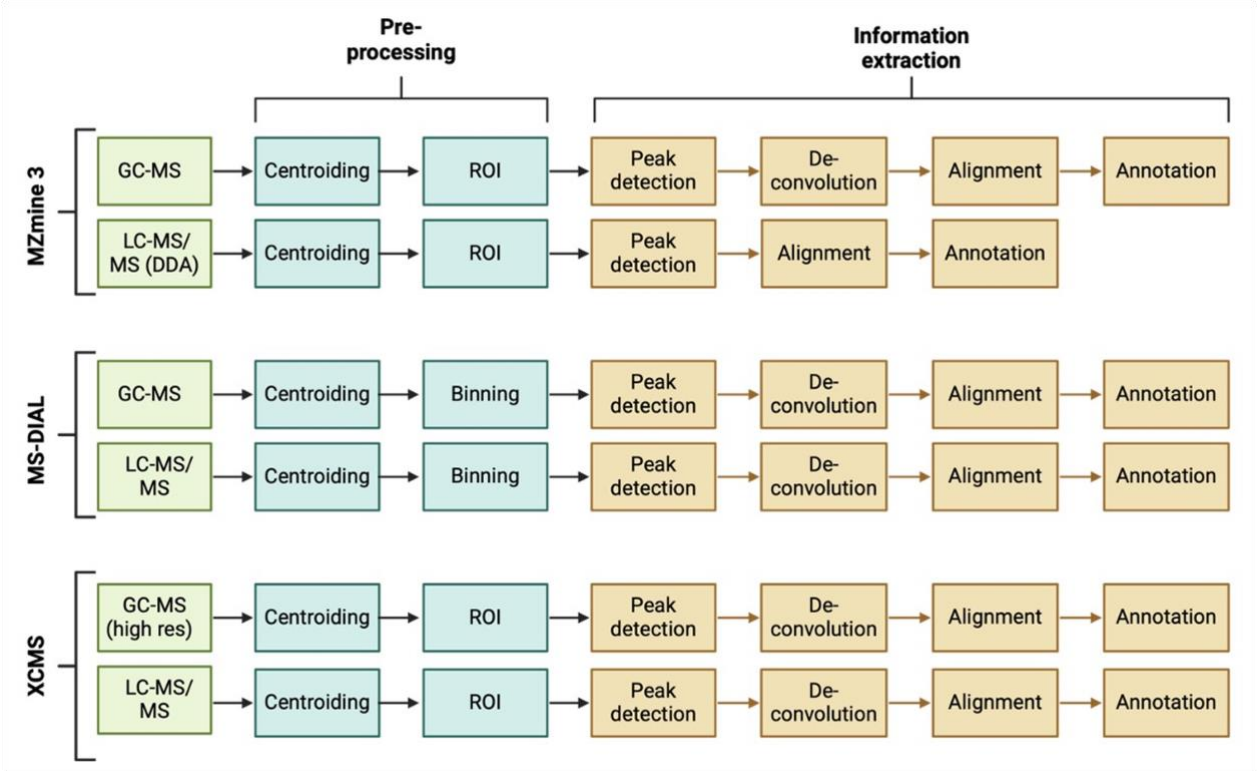


***Figure 24**: Processing of GC-MS and LC-MS/MS data in feature-based workflows implemented in MZmine, MS-DIAL, and XCMS. All workflows follow a similar structure but offer different data processing capabilities and have implemented different functions to extract chemical information from chromatographic raw data. The functions used for information extraction will be investigated in the next Sections. Created with BioRender.com*

5.1.1 Peak detection

In all investigated software packages, peak detection is the first data processing step following the raw data pre-processing steps described in Section 3.4.4. The goal is to identify signals in the raw data that represent chemical information while discarding noise. During the peak picking step, the apex position and peak boundaries in the retention time domain are determined. This information is crucial for subsequent data analysis steps, such as deconvolution or alignment. If a peak picking algorithm fails to detect true peaks, there is a high risk that chemical information is lost. Conversely, if the algorithm identifies many false positive peaks that do not contain chemical information, these may lead to spurious results and incorrect identifications if not filtered out in later processing steps.

The fact that all reviewed software packages offer "gap filling" as a post processing step to fix errors made in the peak detection indicates that this first step is challenging.[35,325,326]

Table 1 provides an overview over the algorithm used for peak detection in the respective software packages, provides the number of user-defined parameters, and references that describe the functionalities and the performance tested in benchmark studies.

***Table 1**: Summary of the peak picking algorithms implemented in MZmine 3, XCMS, and MS-DIAL.*

	Algorithm	**Parameters**	**Functionality**	**Performance**
MZmine	modified *centWav*e	6 (+4)*	229,327,328	43,329
XCMS	*centWave*	6 (+4)*	43,230	43,329
MS-DIAL	Derivative	1 (+4)*	34,330,331	289

*(+4) parameters refer to the EIC construction step, which precedes the peak detection.

The peak detection procedure implemented in MZmine and XCMS are conceptually similar, as both software packages use variants of the *centWave* algorithm.[230,327] The principle of the *centWave* algorithm is that it applies a matched filtering procedure at different scales using the continuous wavelet transform.[230,328] The kernel function used in *centWave* is a "Mexican hat", which is a mirrored version of the second derivative of a Gaussian. Hence, using the continuous wavelet transform offers the advantage of matching filters of different widths, which improves the identification of peaks having different widths. In this procedure, peaks are detected as local maxima in the wavelet coefficients at different scales (different widths of the matched filter). More precisely, a "valid" peak is identified if the positions of the maxima found at different wavelet scales are within a defined retention time tolerance. Connecting the maxima at different scales that are within the retention time tolerance gives a "ridge line".[328]

Myers et al. outlined some problems with the implementation of the *centWave* algorithm in Bioconductor[332] that was used in the past by XCMS and MZmine.[43] These problems have been addressed in the *centWave* algorithm in the ADAP workflow which is implemented in MZmine.[327] One of the issues outlined and corrected by Myers et al. is that the *centWave* algorithm used by XCMS requires no minimum length of the ridge line as opposed to the procedure described by Du et al..[43,327,328] This leads to a high number of false positives, whereas other issues detailed by Myers et al. increase the number of false negatives.[43] Both, MZmine and XCMS, combine peak detection with multiple filtering steps to reduce the number of false positives. For instance, intensity values are compared to a user-defined signal-to-noise threshold to only keep potential peaks with a sufficient number of scans above the threshold.[43] Conversely, MS-DIAL uses a relatively simple algorithm to determine peaks based on noise thresholding and differential calculus. The procedure is shown in a simplified form in **Figure 25**, which has been inspired by the visualization given in the supporting information of Tsugawa et al.[34] The peak detection algorithm starts, by smoothing the raw data using a linearly-weighted moving average (this step has been omitted in **Figure 25**).[331] The window width for this smoothing operation is the only parameter of the peak that needs to be defined by the user.[34] Afterwards, the absolute amplitude distance between consecutive scans, as well as the first and second derivative of a constructed EIC are calculated. The derivatives are calculated using the Savitsky-Golay procedure, as well, but the window size in this step is fixed to five data points.[331] Based on the calculated signal alterations, three different filters called amplitude filter (AF), first derivative filter (FF) and second derivative filter (SF) are calculated (exact procedure can be found in Tsugawa et al.[34]). A peak is detected, if the second derivative of the signal is below SF and the first derivative changes its sign (dashed black line in **Figure 25**). Left and right peak edges are found as the second consecutive point in the interval for which the amplitude distance and the first derivative are above AF and FF, respectively (starting from the left). On the right side the procedure is analogous with the distinction that the first derivative needs to be below FF. The peak boundaries are found as the local minima in a ± 5 point interval around the peak edges.

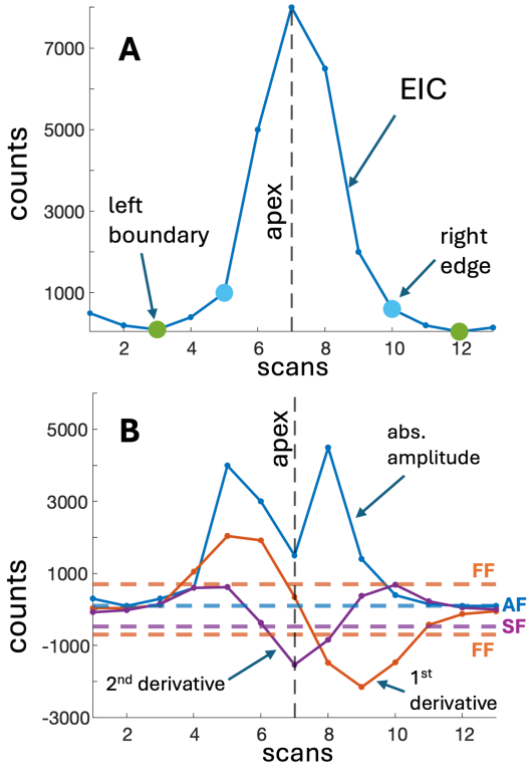


***Figure 25**: Visualization of the peak picking procedure implemented in MS-DIAL. **A:** Smoothed raw data (EIC) is shown. The black dashed line is showing the peak position, and blue and green dots are showing the peak edges and boundaries, respectively. **B:** Amplitude-distance (blue, solid), first derivative (orange, solid), and second derivative (purple, solid), and the respective thresholds (same color-coding but dashed lines) are shown. Peak position is found by identifying the point in which the first derivative is changing its sign, and the second derivative has its minima below SF. For determination of peak edges and boundaries see description in the main text.*

In a recent benchmarking-study, Guo et al. compared different peak picking algorithms including the ones implemented in MZmine, XCMS and MS-DIAL.[329] The results of the study showed that the peak picking implemented in MS-DIAL yields significantly lower false positive and false negative rates, compared to the *centWave*-based algorithms implemented in MZmine and XCMS.[329] Furthermore, the "adjusted *centWave*" algorithm implemented in MZmine performed significantly worse than the "original *centWave*" algorithm implemented in XCMS.[329] The latter finding is surprising, since the modifications made in MZmine's *centWave* version were based on a detailed comparison between the implementations in XCMS and MZmine.[43,327] However, this result may indicate that due to the complex implementation (involving five consecutive filters) and a large number of parameters, even rational changes in individual steps of the peak picking algorithms can lead to unexpected outcomes. The mechanistic investigation by Guo et al. into differences in the peak picking algorithms also led to the hypothesis that the ROI binning implemented in XCMS and MZmine may be responsible

for the observed performance gap compared to MS-DIAL, which uses a simpler equidistant binning.[329]

5.1.2 Deconvolution

Deconvolution, discussed in detail earlier (see Sections 3.2.4 and 4.1), separates overlapped analyte signals in the chromatographic and mass spectral dimensions to extract and group analyte specific EICs, from which mass spectra and relative concentrations can be derived (see **Figure 15**). This process is also referred to as componentization.[97,198]

Table 2 provides an overview of the algorithms used for deconvolution in MZmine, XCMS, and MS-DIAL, along with their applicability to different data types. Additionally, it lists the number of user-defined parameters and references that describe the functionalities. No study that systematically compared the performance of these deconvolution algorithms was found in the literature.

***Table 2**: Summary of the deconvolution algorithms implemented in MZmine, XCMS, and MS-DIAL.*

	Algorithm	**Data type**	**Parameters**	**Functionality**
MZmine	ADAP 4.0 (MCR)	GC-EI-MS	4	35,258
XCMS	CAMERA	LC-MS	2	261,333
MS-DIAL	*MS1Dec*	GC-EI-MS	1*	259,325,334
	MS2Dec	LC-MS/MS	1*	34,259,325
	CorrDec	LC-MS/MS	3	335

*Implementations of comparable algorithms in ADAP 3.2[336] and AMDIS[259] have 10 parameter and therefore it is assumed that many parameters have been fixated with best guesses by the developer.

First, it is important to recognize that the software packages have different deconvolution capabilities. While MZmine and MS-DIAL can process GC-EI-MS data, CAMERA has been developed for annotation and componentization of LC-MS data, though workflows for analyzing high resolution GC-MS data with CAMERA have been described.[337] On the other hand, MZmine does not offer deconvolution tools for LC-MS data in their workflows. MS-DIAL has the broadest capabilities in terms of deconvolution, offering two functions for deconvolution of LC-MS data and one function for GC-EI-MS data.

Since MCR has been comprehensively discussed in Chapter 4, it will not be detailed here. However, a few aspects with regards to the implementation in MZmine are worth mentioning. At first, MZmine uses a hierarchical approach[257], fitting multiple MCR models to non-

overlapping retention time windows that are identified by hierarchical cluster analysis (see Section 4.1). The maximum width of the retention time window is a user-defined parameter. The number of components for the MCR model is estimated by univariate clustering on the peak apex retention times obtained from the earlier peak detection. Optionally, the EIC profiles can be smoothed to obtain a clearer grouping based on the retention times. Two parameters need to be defined for the univariate clustering, the retention time tolerance defining the cluster borders, and the minimum number of peaks defining a cluster. Another aspect is that the MCR version implemented in MZmine is not leveraging information across samples, e.g., by using the extended MCR approach (see Section 4.2.2).[262]

The CAMERA algorithm has originally been proposed as an annotation tool for LC-MS data and provides only limited deconvolution capabilities. **Figure 26** visualizes the different steps of the algorithm, which will be explained in the following. In a first step, the algorithm finds the peak with the highest intensity in the feature list created in the peak picking step and defines a small retention time window around the peak maximum. A pseudo spectrum is created from all features within this retention time window (see **Figure 26A**).

In a second step, m/z differences between all EICs in the pseudo spectrum are calculated and isotope peaks, such as those related to natural abundance of C^{13}, are annotated (**Figure 26B**). Afterward, pairwise score values are calculated based on the similarity of the EICs and the presence of an isotope relation. If a set of multiple samples is analyzed, the Pearson correlation between the intensity ratios of the EIC pairs across the samples is included in the score calculation and used as a cut-off criterion.[261] However, in this scenario, retention time shifts across samples need to be corrected beforehand using alignment algorithms (see Section 5.1.3). Based on the calculated score matrix (containing the pairwise score values), a graph-based clustering is used to separate the EICs from the pseudo spectra into the respective component spectra (as shown in **Figure 26C**).[261] Finally, adduct fragments are annotated using a rule-based procedure (**Figure 26 D**). The algorithm then returns to the first step to find the next highest peak among the remaining EICs in the list. The user-defined parameters required by CAMERA are the correlation threshold for the EIC similarity within a sample and the correlation threshold for the intensity comparison across samples.[333]

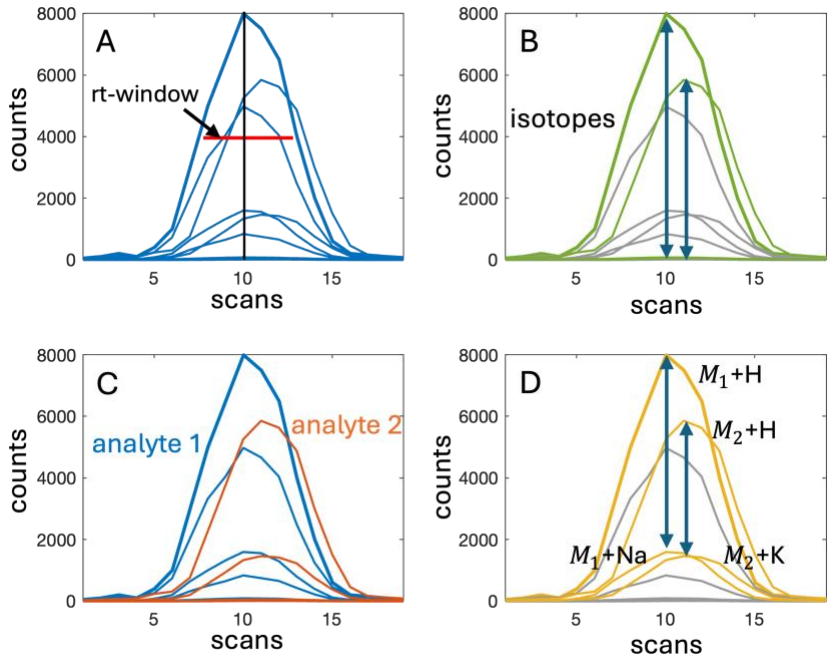


***Figure 26**: Schematic visualization of the different steps of the CAMERA procedure. **A**: Grouping EICs in retention time window around the apex of highest intensity peak to create a pseudo spectrum. **B**: Assigning isotope relations based on m/z differences. **C**: Grouping of EICs into component spectra using graph- based clustering on pairwise calculated similarity scores (see main text). **D**: Rule based annotation of adducts.*

The actual deconvolution in CAMERA is the graph-based clustering, which is conceptually similar to the procedure that was implemented in ADAP 1.0 for the deconvolution of GC-MS data.[338] The problem with this approach is that it cannot separate co-eluting isobaric ions, meaning signals that overlap in both the retention time and the mass spectral domains.[339] These co-eluting isobaric ions will be assigned to either one of the component spectra and will be missing in the other one. However, since soft ionization LC-ESI-MS produces less in source fragments than GC-MS, co-eluting, isobaric ions are less likely to occur, and grouping similar EICs might be a sufficient componentization strategy.[312] The situation differs for LC-MS/MS data measured in DIA, especially in all ion fragmentation experiments, as discussed in Section 3.4.[34,198]

Of the software packages compared, only MS-DIAL offers deconvolution for LC-MS/MS DIA data. The *MS1Dec* and *MS2Dec* algorithms, implemented in MS-DIAL are similar to the deconvolution algorithm used in AMDIS[259] for deconvolution of GC-EI-MS data. The difference is that the *MS2Dec* algorithm works on tandem mass spectrometry data, using peaks detected in MS^1 data to define retention time windows for deconvolution of the MS^2 data. Since the details of the AMDIS, *MS1Dec*, and *MS2Dec* algorithms are cumbersome to describe and

can be found in the literature[34,259,340,341], the following explanation will focus on the conceptual idea of the *MS2Dec* algorithm.

The *MS2Dec* algorithm works on individual samples, successively deconvoluting selected retention time windows. In a first step, a retention time window is defined around a given peak detected in MS^1, and the MS^2 data within these boundaries is extracted (**Figure 27A**). In the next step, all EICs extracted from MS^2 are smoothed and baseline-corrected (step not shown in **Figure 27**). Model peaks are then extracted for each of the co-eluting species using the procedure described by Stein and Hiller et al.[259,340] Briefly, two criteria are assessed for each peak, called *sharpness* and *ideal slope*. Peaks scoring highest on these two criteria are selected as model peaks (usually, high abundant peaks score highest since their shape is less affected by noise as compared to low abundant peaks). In **Figure 27C**, the detected peaks in MS^2 are shown, with the respective model peaks overlaid as colored lines. Finally, all EICs with detected peaks are decomposed using the least squares procedure. To explain the procedure, let $\mathbf{c}_j$ $(I \times 1)$ be a selected EIC in the retention time window, with I being the number of retention time scans in the selected window. Then the deconvolution procedure finds the linear combination, shown in Eq. 22, best approximating $\mathbf{c}_j$. In Eq. 22, $\mathbf{m}_{r=1,2,3}$ are the model peaks, $\mathbf{n}$ is a $(I \times 1)$ vector holding the scan numbers, $\mathbf{k}$ is a $(I \times 1)$ vector of ones, and a, b, c, d, and e are the coefficients to be determined using the least squares procedure.

Equation 22: $$\mathbf{c}_j = a\mathbf{m}_1 + b\mathbf{m}_2 + c\mathbf{m}_3 + d\mathbf{n} + e\mathbf{k}$$

Hence, the *MS2Dec* algorithm in MS-DIAL can resolve in each retention time window three co-eluting peaks, a linear baseline (modeled by the term $d\mathbf{n}$) and a constant offset (modeled by the term $e\mathbf{k}$), as described by Tsugawa et al.[34] In **Figure 27D** is an example shown of an EIC that is the composite of two co-eluting peaks (black line).

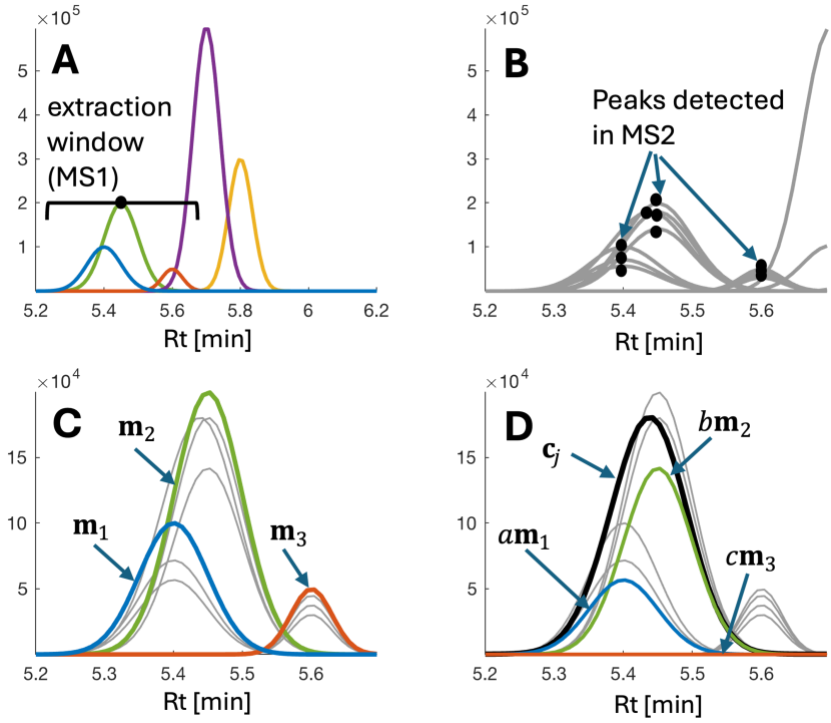


***Figure 27**: Visualization of the deconvolution steps of the MS2Dec algorithm. **A**: Selection of the retention time window around the green peak detected in MS^1. **B**: EICs within the selected retention time window extracted from MS^2. Peak picking is performed after smoothing and baseline correction of the EICs (steps not shown). **C**: Model peaks that have been selected based on different peak shape criteria (see main text or literature[259,340]). **D**: Example of the deconvolution of one composite EIC (black line) using the three model peaks shown in **C**, and Eq. 22. Only a and b are non-zero coefficients.*

The least squares procedure defined by Eq. 22 can resolve the shapes of the co-eluting blue and green EICs by finding non-zero coefficients for a and b. All other coefficients (c, d, and e) are zero. The critical step of the *MS1Dec* / *MS2Dec* algorithm is the selection of good model peaks. For instance, Smirnov et al. describe scenarios where the model peak selection can fail.[258] Their conclusion was that for GC-EI-MS data, MCR is a more robust deconvolution method, yielding higher quality of the extracted spectra.[258] It is surprising, that the *MS1Dec* / *MS2Dec* algorithms in MS-DIAL require only one user-defined parameter[258,325], whereas similar algorithms in AMDIS or ADAP 3.2 require selection of up to ten parameters.[26,336] It is speculated that in *MS1Dec* / *MS2Dec* many parameters have been set to fixed values, like the threshold for the *ideal slope* criterion: "In our software program, an ideal slope score of >0.95 is necessary to be considered a 'model peak'".[34]

The other algorithm implemented in MS-DIAL for deconvolution of tandem mass spectrometry data is the *CorrDec* algorithm.[335] Unlike the *MS2Dec* algorithm, *CorrDec* works on sets of samples by correlating the abundances of precursor ions in MS1 with the abundances of product ions in MS2. Product ions with a high correlation to their respective precursor ion are selected to construct the MS^2 component spectra, which can then be compared to a database for

compound annotation. One drawback of this procedure might be that relying solely on the variation of analyte abundances across samples (without considering EIC shape information) may not be very discriminative. As a result, the quality of the extracted MS^2 spectrum may suffer from many false positive product ions. Furthermore, the procedure requires aligned MS^1 peak lists, meaning that retention time shifts occurring across samples need to be corrected beforehand.

In summary, the *CorrDec* algorithm is a similarity-based grouping algorithm rather than a deconvolution algorithm with the similar limitations as the CAMERA approach discussed above.

5.1.3 Alignment

Due to the lack of control over all variables influencing the chromatographic process, the retention times at which analytes elute will vary across samples and even across replicated measurements of the same sample (see Section 3.2.1). Additionally, m/z values of precursor and product ions of the same analyte may vary slightly across samples, because of instrumental fluctuations or raw data pre-processing steps (see Section 3.4.4). However, to allow for a rigorous statistical analysis of the analytical results, it is crucial to establish correspondence between the detected and deconvoluted peaks across samples.[342] Many different alignment algorithms have been proposed to establish correspondence between peaks in chromatographic data, 50 of which have been reviewed by Smith et al.[343]

Table 3 provides an overview over the algorithms implemented in MZmine, XCMS, and MS-DIAL used to establish correspondence between samples. Furthermore, the number of parameters that need to be defined by the user, and references describing the functionalities are provided. No study systematically comparing the performance of these alignment algorithms was found in the literature.

***Table 3**: Summary of the alignment algorithms implemented in MZmine, XCMS, and MS-DIAL.*

	Algorithm	**Parameters**	**Reference**
MZmine 3	*Join Aligner*	4	[35,344]
	RANSSAC	2	[345,346]
XCMS	*Obi-Warp*	6	[326,347]
MS-DIAL	*Join Aligner*	4	[34,325]

The alignment step is usually performed after deconvolution; however, some deconvolution algorithms require aligned peak lists as input, such as the *CorrDec* algorithm in MS-DIAL. If LC-MS data has been processed using MZmine, alignment is performed directly after the peak picking without prior deconvolution. In the case of GC-MS data, MZmine does not use the *Join Aligner* (which will be explained in the following) but matches peaks across samples based on spectral similarity of the deconvoluted mass spectra.[229]

The simplest strategy to establish correspondence is by performing a binning in mass and retention time direction, for instance by assigning all peaks within a specified (m/z × retention time)-window to the same analyte. The *Join Aligner* implemented in MZmine is an example of such an algorithm.[231] A slightly modified version of MZmine's *Join Aligner* is used in the MS-DIAL software.[34] In MZmine 2 an alternative alignment algorithm called RANSAC (random sample consensus) was proposed.[345,346] However, in a recently published online protocol the use of the *Join Aligner* is recommended and therefore the RANSAC algorithm will not be further discussed herein.[35] Unfortunately, it was not straightforward to find a detailed description regarding the implementation of the *Join Aligner* in MZmine. Moreover, different scoring functions have been documented in the online protocol[35] and the official online documentation of MZmine[348], which has been referenced in the online protocol for more detailed explanations. Additionally, no reference that points to an article clearly describing the algorithmic procedure of the *Join Aligner* has been provided in the initial MS-DIAL paper by Tsugawa et al.[34] However, based on the visual description in the Supporting Information of Tsugawa et al.[34] and based on the detailed description in Katajamaa et al.[344] it is possible to decipher how the *Join Aligner* in MZmine and MS-DIAL works.

The algorithmic procedure starts by initializing a reference list as the peak list of a user-selected sample. Every peak in the list is characterized by a retention time and an m/z value. Hence, the reference list has initially two columns, retention time and m/z value, and in the beginning as many rows as peaks detected in the user-selected sample. In the next step the reference list is expanded by sequentially adding peaks from all the other samples under the condition that a new peak is not already represented by any other peak in the reference list. To check if a new peak is already represented in the list, m/z differences (Δ_{mz}) and retention time differences (Δ_{rt}) are calculated between the candidate peak (to be added to the list) and all peaks already in the list. If the calculated m/z and retention time differences are larger than the user-defined thresholds T_{mz} and T_{rt}, the candidate peak is added. Once all "unique" peaks from all samples have been added to the reference list, score values are calculated for every peak in every sample with every peak in the reference list (according to Eq. 23 or 24). A peak detected in a sample is

aligned to the peak in the reference list that yields the highest score (meaning it has the lowest deviation in m/z and retention time). The contribution of deviations in m/z and in retention time can be modified by adjusting the weights w_{mz} and w_{rt} in the scoring function shown in Equation 23 and Equation 24. In MZmine the scoring function shown in Equation 23 is implemented and in MS-DIAL the scoring function shown in Equation 24 is implemented.

Equation 23: $$score = \left(1 - \frac{\Delta_{mz}}{T_{mz}}\right) * w_{mz} + \left(1 - \frac{\Delta_{rt}}{T_{rt}}\right) * w_{rt}$$

Equation 24: $$score = w_{mz} * e^{-0.5\left(\frac{\Delta_{mz}}{T_{mz}}\right)^2} + w_{RT} * e^{-0.5\left(\frac{\Delta_{RT}}{T_{RT}}\right)^2}$$

Some limitations associated with the *Join Aligner* procedure will be discussed below. First, the exact position and m/z value of the "unique peaks" in the reference list depend on the sample used for the initialization of the reference list and the order in which the samples are compared to the reference list. Additionally, *Join Aligner* is a greedy algorithm, which means that it finds in each step the best match of a peak in the sample, to a peak in the reference list, without considering if there is a "better-matching" peak in the sample. This is problematic because the alignment results depend on the order in which the sample peaks are matched with the reference list. Finally, the large number of parameters can make the results biased toward the subjective choices of the user.

A conceptually different approach for alignment is represented by the *Obi-Warp* algorithm implemented in XCMS.[347] The *Obi-Warp* algorithm is a constrained version of the dynamic time warping algorithm (DTW), which uses dynamic programming to achieve a globally optimal alignment (as opposed to the *Join Aligner*).[349,350] The procedure of the DTW algorithm is exemplified in **Figure 28**. In DTW, shift is corrected by mapping the time axes of a sample (orange line in **Figure 28A**) and a reference chromatogram (blue line in 28A) to a common time axis.[351] This correspondence mapping is also called warping path and is visualized as dashed black line in **Figure 28B**. A one-to-one correspondence between two identical (non-shifted) samples would yield the solid black line in **Figure 28C**. The result of the alignment is shown in **Figure 28C**.

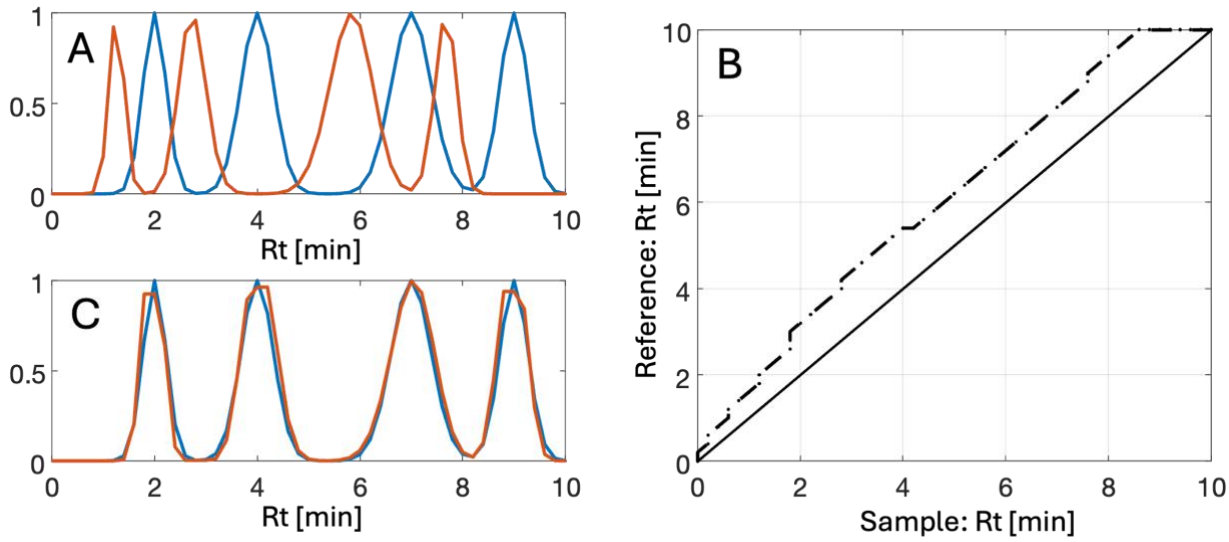


***Figure 28**: Visualization of the alignment of two chromatographic profiles using dynamic time warping (DTW). **A**: Reference chromatogram (blue) and a shifted sample chromatogram (orange) are shown. **B**: Warping path for the alignment of the sample chromatogram to the reference chromatogram (black, dashed). The warping path of two identical, unshifted samples is indicated by the solid black line. **C**: Result after alignment. Peak shapes of the sample chromatograms have changed, as compared to **A**.*

Finding the optimal warping path $F = (t_{Ref}(k), t_{sample}(k))$ with $k \in \{1, \dots, K\}$ and K being the length of the warping path involves optimizing a cost function $D(F)$, which measures the similarity or dissimilarity between the reference and the sample chromatogram after warping.[351] Tomasi et al. showed that the quality of the alignment obtained with DTW is highly dependent on the parametrization of the algorithm and the implementation of suitable constraints.[351] In their study, unconstrained DTW led to distorted peak shapes, which can be emphasized by comparing, for instance, the first peak of the orange curves in **Figure 28A** and **Figure 28C**. The *Obi-Warp* algorithm does not use a slope constraint, as recommended by Tomasi et al., but instead implemented a spline-based smoothing to enforce bijectivity and limit the distortion of peak shapes and areas.[347,351] Another limitation related to warping algorithms in general is that the results depend on the selected reference sample.[343] Skov et al. have provided a procedure to select suitable reference chromatograms and automate the parameterization for the correlation optimized warping (COW, which can be seen as a special case of constrained DTW).[351,352] However, Skov et al. argue that applying warping-based alignment to chromatographic data in which peak shapes change across samples will lead to incorrect results.[352] Further, warping-based alignment methods assume that the retention order is preserved, which is a shortcoming noted by Prince et al.[343]

5.1.4 Annotation

Annotation describes the process of assigning chemical identities to chromatographic peaks using the extracted mass spectral and retention time information. According to the Metabolomics Standard Initiative (MSI), it is required to verify the identity of an metabolite with at least two orthogonal pieces of information, such as exact mass, retention time, or fragment spectra.[96] For GC-MS the combination of retention indices[353,354] and fragment spectra and for LC-MS/MS the combination of the exact mass and fragment spectra are commonly used to annotate structures.[69,170]

However, in 2015, da Silva and co-workers stated that only 1.8 % of mass spectra collected in untargeted studies are annotated[355] highlighting that annotation presents the major bottleneck for progress in metabolomics.[67,356] The annotation problem can be distinguished into the problems of identifying "unknown knowns" (where a reference entry exists in a mass spectral database) and "unknown unknowns" (where no reference entry exists in a mass spectral database).[67] Enormous efforts have been made in recent years to facilitate the identification of unknown knowns[38,170,357] and unknown unknowns [358–362].

Open mass spectral databases and data sharing initiatives are essential for the identification of unknown knowns.[170,357] Due to the high mass spectral reproducibility, databases based on 70 eV GC-EI-MS have a long history and large commercial databases like the NIST-DB[363] have been built. Open-source databases like the GOLM-DB[364] also exist. On the other hand, databases for LC-MS/MS are relatively immature and suffer from lacking standardization of experimental conditions for mass spectra acquisition,[182] which is illustrated by the example given in **Figure 29**. In **Figure 29** mass spectra of *N,N-Diethyl-meta-toluamide (DEET)* are shown, which have been gathered from Massbank Europe (https://massbank.eu, 21.07.2024). The mass spectra have been recorded at different collision energies (30 eV, 40 eV, and 50 eV), on different LC-ESI-QTOF instruments (TripleTOF 5600 Sciex and Bruker maXis Impact), and under different chromatographic conditions (A: Water 0.1% Formic acid, B: Acetonitrile 0.1% Formic acid vs. A: 90:10 water:methanol, 0.01% formic acid, 5mM ammonium formate B: methanol, 0.01% formic acid, 5mM ammonium formate). Despite being mass spectra of the same chemical compound, significant differences in the cosine similarity of the mass spectra can be observed.

Nevertheless, the growth of LC-MS(/MS) databases in the past decade is astonishing. Bittremieux et al. report that the open-source global natural products social molecular network (GNPS) community spectral libraries have grown from 23.790 to 586,647 MS/MS spectra in the period from 2014 to 2022.[357] Furthermore, open-access search tools that integrate multiple

databases are of great importance. For instance, the MASST tool[105] allows querying mass spectra against multiple mass spectral databases including all MassBanks[365] and the GNPS community spectral libraries.[105,366]

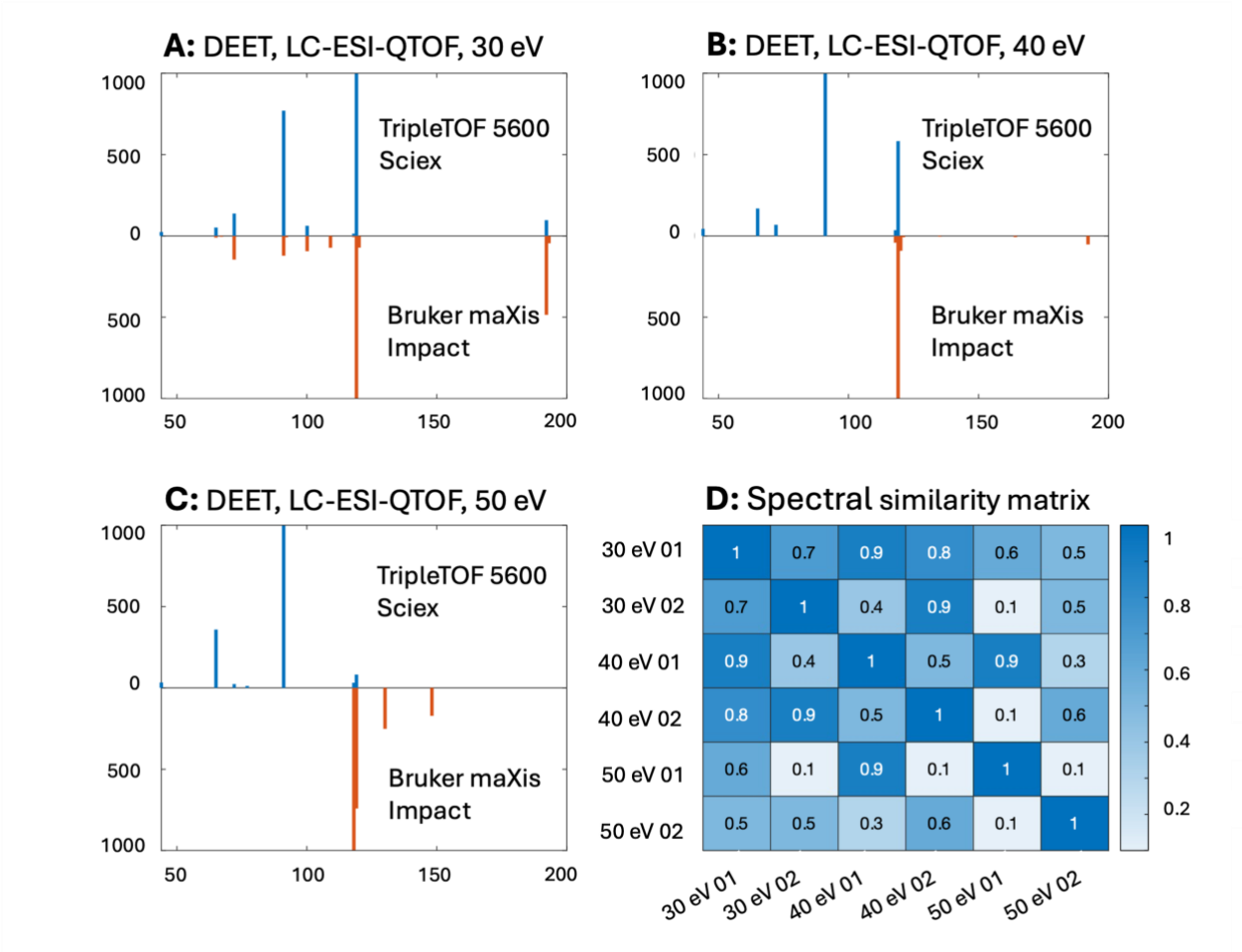


***Figure 29**: Mass spectra of N,N-Diethyl-meta-toluamide, recorded at different collision energies and on different instruments. **A:** Mass spectra recorded at 30 eV. Both instruments produce the same product ions, relative abundances vary considerably. **B:** Mass spectra recorded at 40 eV. Larger differences in the fragmentation pattern, no product ions with m/z smaller than 118 have been extracted from the Bruker maXis Impact raw data. **C:** Mass spectra recorded at 50 eV. Mass spectra deviate significantly in the extracted product ions and their abundance. **D:** Similarity matrix showing the cosine similarity between mass spectra. The similarity between the mass spectra of the same compound recorded at the same collision energy but on different instruments and under different chromatographic conditions can vary between 0.7 (30 eV) and 0.1 (50 eV). Similarity of mass spectra recorded on the same instrument and under the same chromatographic conditions can still vary considerably if different collision energies have been used.*

An active field of research is the development of new similarity metrics or scoring algorithms to obtain structural proposals of chemically similar compounds from a database search, even if the true compound is not present in the database.[367] A graph-based approach is, for instance, employed in GNPS, one of the largest annotation tools used today.[361,368] The hypothesis underlying this research is that (chemical) structural similarity (similar functional groups, connectivity, topology)[369] of two compounds can be inferred from their mass spectral similarity.

This assumption is also foundational for *in silico* annotation tools like SIRIUS[359], which use combinatorial approaches and machine learning to predict chemical fingerprints (computerized representations of chemical molecules) from mass spectra. Tools like SIRIUS potentially extend the space of what is "known" from spectral databases to the much larger structural databases like PubChem.[370,371] Furthermore, quantum chemical[372] and machine learning approaches[373] have been used to create large *in silico* spectral databases based on predicted mass spectra.[38]

One great strength of MZmine, MS-DIAL, and XCMS is that they are all interfacing with comprehensive annotation tools like SIRIUS or GNPS and with mass spectral databases like METLIN.[34,35,38,359,361,374–377] However, to effectively leverage this potential, clean and accurate mass spectra need to be extracted from the raw data. Otherwise, compound identification can still be a challenge [Paper IV] and incorrect structure assignments may lead to wrong mechanistic explanations and conclusions.[98]

5.2 Pixel based and tile-based data analysis for comprehensive two-dimensional gas chromatography coupled to mass spectrometry

The information extraction approach introduced in Section 4.3 is also referred to in the literature as peak table-based data analysis.[378] This is because curve resolution or tensor decomposition methods can potentially be used to comprehensively extract all chemical information in the form of analyte mass spectra and peak areas (or peak volumes in the case of two-dimensional chromatography) from the raw data.[300,301] However, due to the computational complexity of the peak table-based data analysis [378,379], the development of pixel-based and tile-based approaches as prioritization and information extraction tools has been focused on. [378,380] Prioritization in this context means, that pixel- and tile-based approaches help identifying chromatographic regions that contain relevant signal, e.g., to describe chemical differences between classes of samples.

5.2.1 Pixel-based approach

The pixel-based data analysis approach is straightforward in that it is a point-to-point (or pixel-to-pixel) comparison between measurements. In this context, every data point in a single chromatographic measurement is considered a pixel.

Considering GC×GC measurements with univariate detection like a FID, a pixel $x_{k=1}^{ij}$ is the intensity value at retention time i in ^{1}D and at retention time j in ^{2}D in the first out of K samples.

This intensity value can now be compared to the intensity values $x^{ij}_{k=2,\dots,K}$ in the other samples at the same position ($i_{k=1} = i_{k=2} = \cdots = i_{k=K}$ and $j_{k=1} = j_{k=2} = \cdots = j_{k=K}$). The objective of the pixel-based method is to find pixels across samples that contain relevant information. One approach to achieve this is the calculation of pixel-wise Fisher ratios.[380] This method requires grouping of the samples into classes, which are often known *a priori* from the experimental design (untargeted supervised studies).[380] Alternatively, pooled quality control (QC) samples or blank samples can be used to construct a class against which the samples can be compared. The Fisher ratio F_{rat} for a given pixel is calculated according to Eq. 25-27.

Equation 25: $$F_{rat} = \frac{\sigma^2_{class}}{\sigma^2_{error}}$$

Equation 26: $$\sigma^2_{class} = \frac{\sum_{c=1}^{C}(\bar{x}_c - \bar{x})^2 N_c}{C-1}$$

Equation 27: $$\sigma^2_{error} = \frac{\sum_{k=1}^{K}(x_k - \bar{x})^2 \sum_{c=1}^{C}(\bar{x}_c - \bar{x})^2 N_c}{K-C}$$

With σ^2_{class} being the between class variance, σ^2_{error} being the within class variance, $\bar{x}_c$ being the mean of class $c \in \{1, \dots, C\}$, $\bar{x}$ being the mean over all samples, N_c being the number of samples belonging to class c, and x_k being the pixel intensity of measurement $k \in \{1, \dots, K\}$. Once the Fisher ratios at all retention time pairs (i, j) have been calculated, a map of the Fisher ratios can be plotted to visualize which pixels are most informative in the sense that they explain between class variances. If a multichannel detector such as TOFMS is used, pixel-to-pixel relationships can be calculated for each mass channel at a given retention time. Hence, a Fisher ratio map can be plotted for each mass channel to identify and eventually group the informative mass channels. The pixel-based approach, in combination with the Fisher ratio method, has been used as a prioritization tool in comprehensive GC×GC-TOFMS measurements, to find chromatographic regions which were then further analyzed using PARAFAC or PCA.[380–382] Instead of using the Fisher ratio method, the fingerprints of different samples may be compared. In this approach, GC×GC measurements are vectorized to form the matrix $\mathbf{X}_{K,aug}$ $(IJ \times K)$ for a set of K samples, where I denotes the first retention dimension, and J denotes the second retention dimension. Furbo et al. described this strategy, using PCA on $\mathbf{X}_{K,aug}$ to investigate an industrial hydrotreatment process.[379] This idea can conceptually be used for GC×GC with univariate or multichannel detectors, however, vectorizing GC×GC-TOFMS measurements

creates an enormous number of variables (~10.000.000-100.000.000), which will increase the uncertainty in the estimated latent variables.[265] Applications of the pixel-based approach are by no means limited to comprehensive two-dimensional chromatography, but can also be found in GC-FID and GC-MS data analysis in environmental monitoring and metabolomics studies.[383–386]

The pixel-based approach requires comprehensive data pre-processing, involving baseline removal, smoothing, normalization, and scaling, as has been described in detail by Furbo et al.[379] However, it is crucial that retention time shifts in 1D and 2D are corrected to establish exact pixel-to-pixel correspondence across samples. Some alignment methods and their limitations have been discussed in Section 5.1.3, and further methods can be found in the literature.[387–389] For the alignment of GC×GC data, it has been argued, for instance, by Parsons et al. that the alignment of the first retention dimension (1D) is difficult due to the low information density (few modulations per peak).[390] Imperfect alignment is detrimental for pixel-based analysis in two ways, as it reduces the sensitivity for true positives and can lead to the discovery of artifactual false positives.[378]

5.2.2 Tile-based approach

The tile-based approach was developed to overcome the sensitivity of pixel-based methods to correspondence issues.[378,390] The conceptual difference between the tile-based and the pixel-based approach is that two-dimensional bins (tiles) of multiple pixels are compared across sample classes instead of individual pixels.[378,390] Thus, due to the pooling of multiple pixels, the strict requirement for pixel-to-pixel correspondence is relaxed in tile-based methods. The conceptual difference between pixel-based and tile-based approaches is visualized in **Figure 30A**. **Figure 30A** visualizes the TICs of four samples, where the same analyte elutes at different retention times. While the retention time shift introduces variance at the pixel level (compare the red square), the variance at the tile level (indicated as the grey shaded area) remains unaffected. The size of the tiles needs to be adjusted for a specific dataset and is typically set to match the respective ^{1}D and ^{2}D peak width, plus a margin accounting for retention time shift. The rationale is that a tile should be wide enough to fit most peaks within its boundaries, but sufficiently small so that ideally only one peak occurs in each tile. However, there is a high risk that some peaks will not be sampled correctly if only one tile-grid is used, as peaks could be positioned on the borders of two tiles. Therefore, the tile-grid is shifted in different directions to increase the probability that each peak gets at least once sampled sufficiently (see **Figure 30A**).[378,390]

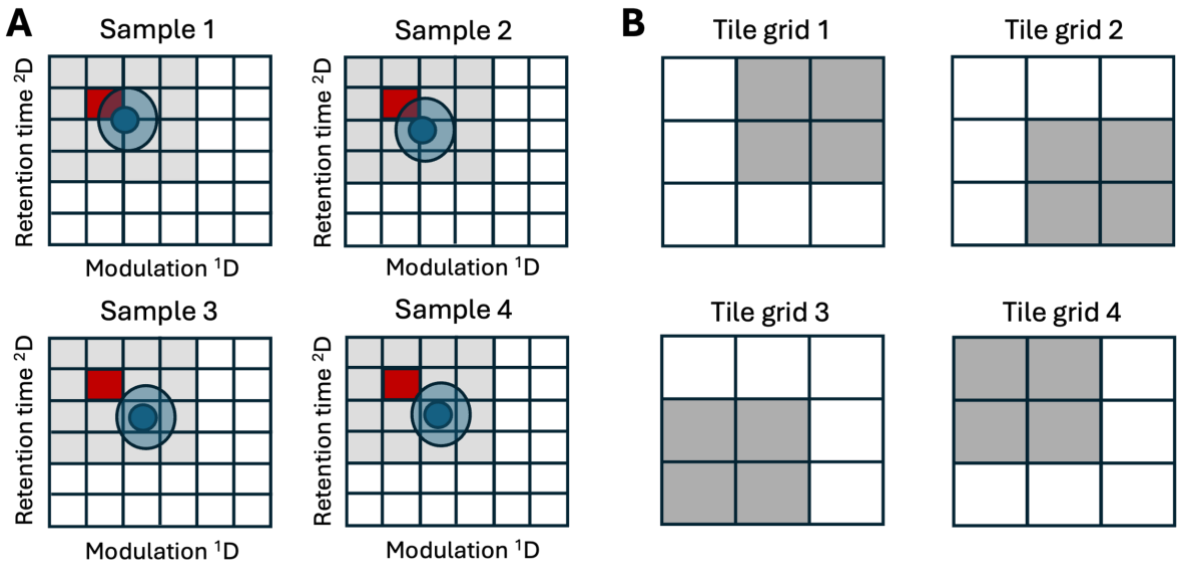


***Figure 30**: Comparison of tile-based and pixel-based approach for the analysis of GC×GC data. **A**: Retention time shifts of the blue peak introduce variance on a pixel-based level as can be seen by comparing the overlap of the blue peak with the red square. Variance on the tile-based level (indicated by the grey-shaded area) remains unaffected by retention time shift. **B**: To reduce the risk of peak splitting, every peak gets sampled at four different grid positions.*

Since every peak is sampled at four different grid positions, a filtering procedure needs to be applied to remove redundancies and keep only the tiles that best capture each peak (detailed description in Supplementary material of Parsons et al.[390]). The tile-based approach in combination with the Fisher ratio method gained significant popularity for the analysis of GC×GC data and was recently implemented into the commercial ChromaTOF® software distributed by LECO.[89]

However, there are some limitations to the tile-based Fisher ratio approach, which have been addressed in recent studies. One limitation is that the Fisher ratio, as defined by Eq. 25, is lacking discriminative power in situations in which the within class variance is not homogeneous, such as when the variance in the "patient" or "treatment" class is substantially larger than the variance in the control class.[391] This situation of unequal variances can lead to a reduced sensitivity of the tile-based Fisher ratio approach for the discovery of biomarkers.[392] To address this problem, Prebihalo et al. and Schöneich et al. have suggested the use of alternative Fisher ratios, termed "Control-Normalized Fisher ratio"[391,392] and "Minimum Variance Optimized Fisher ratio"[393], in combination with the tile-based approach. The idea of a "Control-Normalized Fisher ratio" was originally proposed by Brownie et al. who suggested normalizing the between class variance to the variance of the control group only, instead of using pooled within class variances.[391]

Another limitation of the tile-based Fisher ratio approach, regardless of the type of Fisher ratio used, is that it can only investigate differences between user-defined sample classes. Although this condition can often be met by the experimental design (e.g., in metabolomics studies), an untargeted unsupervised approach might reveal more unexpected patterns in the samples. The variance rank-initiated unsupervised sample indexing method has been developed by Cain et al.[73] and adapted to GC×GC-TOFMS by Sudol et al.[394] to overcome this limitation.

Although the tile-based Fisher ratio approach is an excellent tool for prioritization, it may fail to extract clean compound mass spectra in the presence of co-elution and high background signal.[30] However, two recent studies have proposed strategies to extract clean mass spectra from untargeted supervised GC×GC-TOFMS data using the tile-based approach.[30,72] The class comparison-enabled mass spectrum purification (CCE-MSP) method outlined by Ochoa et al. describes a filtering technique that can, in a two-class scenario, remove interferent signals from an analyte signal to produce a cleaned compound mass spectrum.[30] It has been shown that the CCE-MSP method produces cleaner mass spectra for analytes of low abundance compared to MCR, PARAFAC, and PARAFAC2. However, in more complex situations, the use of MCR-ALS or other deconvolution tools after interferent removal with CCE-MSP has been recommended.[30,72] The results reported in [Paper IV] on the performance of MCR-ALS for the extraction of clean mass spectra from trace-level compounds in LC×LC-HRMS support the findings of Ochoa et al. and Cain et al.[30,72] In [Paper IV], a combination of mass filtering and MCR-ALS was proposed to extract clean mass spectra and improve the identification of trace-level compounds in suspect screening of complex wastewater samples.

5.3 Comparison of data analysis methods and workflows

The capabilities and limitations of curve resolution/tensor decomposition, feature-based workflows, and pixel-based/tile-based methods have been discussed and reviewed in previous Sections. This Section aims to compare and discuss fruitful synergies between the different approaches to overcome some of the outlined limitations. As a reminder, the goal of information extraction was defined in Section 4.1 as converting chromatographic raw data into a peak table containing relative concentrations and chemical identities. It was further described that information extraction using curve resolution and tensor decomposition methods is achieved by decomposing selected regions of the raw data into distinct components. The appropriate number of components needs to be estimated either from the raw data e.g., using SVD[266,267], or by using diagnostic tools.[269–271] From all components, the chemically meaningful ones need to

be selected, as they carry the relative concentrations and compound mass spectra. The extracted mass spectra need to be compared to a database for annotation. The strength of curve resolution and tensor decomposition methods, as opposed to the information extraction algorithms implemented in feature-based workflows, is threefold.

First, curve resolution/tensor decomposition can resolve multiple co-eluting peaks and separate chemical information from low-frequency and high-frequency noise. This capability has been significantly enhanced by the algorithms described in [Paper I] and [Paper III]. For very low abundant signals it may be necessary to implement additional pre-processing steps, as has been suggested in [Paper IV]. Conversely, the *MS1Dec* / *MS2Dec* methods employed in MS-DIAL can only resolve a limited number of co-eluting peaks, and additional pre-processing steps, like baseline correction, are required. The CorrDec algorithm in MS-DIAL and the CAMERA method in XCMS have only limited deconvolution capability, since they only group similar EICs. The mass spectra constructed with these methods will be incorrect in the presence of co-elution. The only feature-based workflow that has implemented MCR (as a representative from the curve resolution / tensor decomposition family) is MZmine; however, it is only used for the decomposition of GC-MS data.

The second advantage of curve resolution/tensor decomposition methods is that they can efficiently handle complex data structures either in multiset arrangements or as higher-order tensors. In this way, additional samples will even improve the results of the decomposition, since noise is averaged over more samples. Moreover, in the case of MCR, rotational ambiguity is reduced if multiple samples are analyzed in a multiset instead of being analyzed individually. Since methods like MCR, PARAFAC2, and SIT/SIST can handle retention time shifts (MCR and SIT/SIST can even accommodate changes in retention order), they provide componentization and alignment across samples in a single step. On the other hand, none of the reviewed feature-based workflows analyzes chromatographic data in "batch mode". Even the MCR version in MZmine is not utilizing the advantages of the multiset approach. Therefore, alignment is a crucial step in all workflows, which can lead to problems due to the limitations of alignment algorithms, as discussed in Section 5.1.3.

The third advantage is that curve resolution/tensor decomposition methods require fewer parameters than feature-based workflows. This is partly a consequence of the second advantage since the alignment step in feature-based workflows requires the user to define at least four parameters (see Table 3). The problem of overparameterization has long been recognized as a serious issue in AMDIS, the first automated data processing software for GC-MS data.[26,395,396] Since AMDIS can be seen as the main source of inspiration for some of the feature-based

workflows today[34,334], many of the problems have carried over. More specifically, results from feature-based workflows have been shown to be highly sensitive to user-defined parameter choices and hence suffer from large numbers of false positives and false negatives in cases of inappropriate parameter selection.[43] The issue is that it is not always clear what constitutes appropriate or inappropriate parameter choices. For instance, this is because different workflows use the same parameters to calculate different quantities. The *Join Aligner* implemented in MS-DIAL and MZmine serve as an example to illustrate why this is a problem (see **Figure 31**). Because different scoring functions are used in MS-DIAL and MZmine (see Eq. 23 and 24), the same parameter choices (e.g., rationally derived from the known mass accuracy and retention time shift) will yield different values of the scoring function. Hence, the results of the alignment can be different in the sense, that a peak in a sample might be aligned to different targets in the reference list (compare discussion in Section 5.1.3). Therefore, an appropriate set of parameters, which yield reproducible results across different workflows cannot be selected based on the fundamental data characteristics but need to be found via hyper-parameter optimization.[325,348]

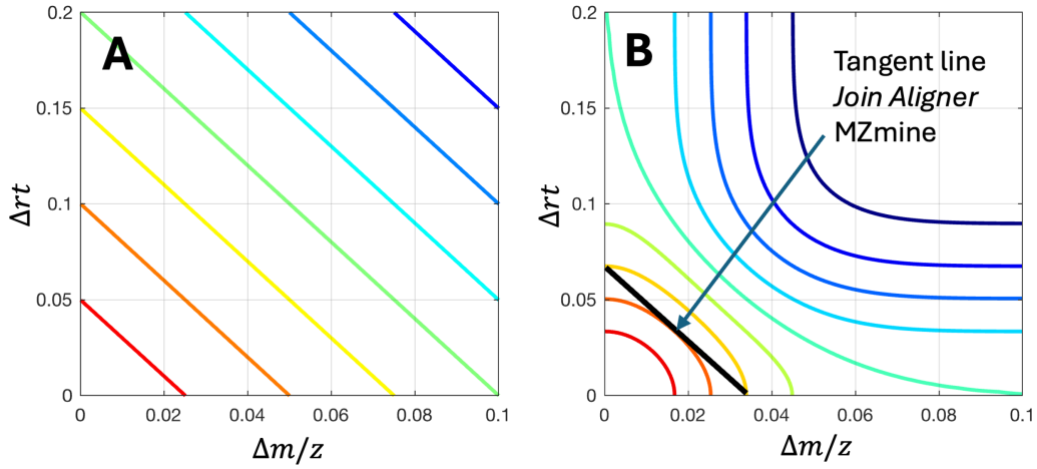


***Figure 31**: Example of the differences in the Join Aligner scoring functions, shown as contour plot. Red contour lines stand for high score values whereas blue contour lines represent low score values. **A**: Score values of the Join Aligner in MZmine. **B**: Score values of the Join Aligner in MS-DIAL. The same set of parameters will yield different results, as can be seen from the black tangent line, representing the MZmine Join Aligner scoring function in **B**. While all combinations of $(\Delta rt, \Delta mz)$ on this tangent line will yield the same score values in MZmine, different score values will be obtained with the Join Aligner in MS-DIAL. Hence, a peak in a sample may be aligned to different targets in the reference list, depending on which Join Aligner version is used, despite using the same set of parameters.*

This lack of transparency is what makes current feature-based workflows essentially black boxes. As a result, many recent studies describe that different feature-based workflows extract different analytical information; however, with varying conclusions regarding the impact on

the downstream data analysis.[39–43] The main conclusion of these studies is that more than one workflow should be used to increase the certainty of the analytical findings e.g., by considering only those peaks as true positives that have been found by multiple workflows.[39,41,42] Of course, this procedure only addresses the problem of false positives, besides being very time consuming. Alternatively, tools for automated hyperparameter optimization have been proposed and shown to improve results significantly.[397–400]

The key difference between chemometric methods like curve resolution/tensor decomposition and feature-based workflows is that chemometric methods have been developed with a strong focus on the chromatographic data structures. This has been emphasized earlier as the abductive approach, using the words of Harald Martens.[238] Thus, the high number of parameters required by feature-based workflows is at least partly due to a lack of structural assumptions. The ability of curve resolution/tensor decomposition methods to help reduce the "overparameterization problem" has also been acknowledged in recent reviews.[82,97] An impressive example of this is the work by Baccolo et al., who developed a fully-automated PARAFAC2-based data analysis workflow for untargeted GC-MS, converting raw data into peak tables without any user-defined parameters.[255] However, the combinatorial, "brute force" approach for the selection of deconvolution windows turned out to be too computationally demanding for routine applications.[255] More sophisticated prioritization methods e.g., based on 1-D versions of the tile-based Fisher ratio may help to make the approach computationally more efficient. Generally, the PARAFAC2-based Deconvolution and Identification System (PARADISe) is a good example of a data analysis workflow that has minimized the need for user-defined parameters.[26,255] For LC-MS data analysis, the ROIMCR method has been implemented into the MCR-ALS GUI.[250,401] Furthermore, curve resolution/tensor decomposition can be flexibly adapted to model more complex multi-block data structures, such as LC-MS/MS DIA all ion fragmentation[198,260], positive and negative ionization[402], LC-UV-MS/MS[403], and likely LC-IM-MS (has not yet been described in literature).

For untargeted GC×GC-MS and LC×LC-MS, only a few dedicated software solutions and even fewer open-source tools exist. Hence, only a small number of studies has been systematically investigating the effect of data analysis tools on the analytical results.[217] Nevertheless, workflows based on the tile-based Fisher ratio have been proven to be versatile for many GC×GC-MS applications.[30,72,73,89,393,394,404] However, the deconvolution capabilities of the tile-based Fisher ratio are limited, and pure component spectra can only be extracted under certain circumstances.[30,72] Given the complexity of GC×GC-MS and LC×LC-MS data, more research effort is needed before comprehensive raw-data-to-peak-table workflows can be

realized. However, testing combinations of tile-based Fisher ratio as prioritization tool together with SIML or MCR for deconvolution could be the first step on this path.[23,161] For low abundant signals, CCE-MSP or the mass filtering approach proposed in [Paper VI] can be applied as pre-processing to increase the SNR and facilitate deconvolution with curve resolution/tensor decomposition methods.[30] However, the SIML method itself has been equipped with noise-filtering capabilities, as described in [Paper III]. It was demonstrated that the additional noise-filtering significantly improves the robustness toward low SNR.

6 Conclusion

This thesis has comprehensively explored the extraction of chemical information from untargeted chromatography hyphenated to mass spectrometry datasets. It has focused on three main areas: first, identifying factors that influence the structure of chromatographic data; second, evaluating the application of curve resolution and tensor decomposition methods for modeling various types of chromatographic data; and third, comparing the efficacy of these methods with other established data analysis techniques.

6.1 Closing remarks on the chromatographic data structure

It can be concluded that many factors are influencing the chromatographic data structure and that artifacts in chromatographic data are the norm rather than the exception – especially in untargeted analysis. Carefully chosen QA/QC procedures such as suitability tests, recovery standards, and pooled QC samples need to be implemented to monitor the fitness of the analytical procedure. It is of utmost importance that the computational methods discussed throughout this work should not be thought of as remedies for correcting bad data. However, since it is impossible to have perfect control over all factors influencing data quality, it is important that chemometric methods can compensate for some artifacts that may be inevitable. In this sense, understanding the chromatographic data structure is crucial for developing chemometric methods that make appropriate assumptions.

6.2 Closing remarks on curve resolution and tensor decomposition methods

It has been discussed that bilinear models are the most flexible models and thus can compensate for artifacts like retention time shifts and peak shape changes across samples. The downside is that bilinear models suffer from rotational ambiguity under the conditions inversely formulated by Manne.[268] The rotational ambiguity translates into uncertainty on the level of analytical results e.g., analyte concentrations and mass spectra.[28] Unique solutions are guaranteed only by trilinear or more generally multilinear models. Unfortunately, chromatographic data often not adheres to the assumption of multilinearity due to the presence of artifacts. Hence, relaxed versions of multilinear models that can handle artifacts like retention time shifts and peak shape changes while providing unique solutions represent the sweet spot to be hit. The methods

proposed in [Paper I-III] may be valuable extensions to the chemometric toolbox in this regard, as they make more relaxed assumptions about the data structure while providing unique solutions. In this sense, the proposed methods complement other methods like PARAFAC2 or the flexible trilinearity constraint.[264,266,288] Nevertheless, examples have been discussed for which even the bilinearity assumption does not hold. Specifically, these examples comprehend matrix effects in LC-MS(/MS), spectral skewing in GC-MS measured on a quadrupole, as well as detector saturation related to ADC/TDC components.[198,201,210] These examples should be further investigated for their practical relevance and may require the development of new chemometric approaches.

6.3 Future perspectives

It is hoped that this work demonstrated that there is not a perfect method or strategy for extracting information from chromatographic hyphenated mass spectrometry data. Instead, a variety of methods exist that are more or less useful, depending on the specific characteristics of the investigated data. Nevertheless, it is believed that chemometric methods can help to reduce the "overparameterization problem" of existing data analysis workflows. To achieve this, chemometrics methods need to be implemented, ideally in open-source workflows and in user-friendly software. This is something the chemometrics community can learn from the success and the popularity of tools like MZmine and XCMS. Furthermore, a higher level of automation in data analysis may also be achieved by incorporating AI into chromatographic data analysis workflows. In a pessimistic scenario, AI could be used to develop the next generation of "black box" information extraction workflows that could suffer from poor generalizability. This risk is increased if improper data augmentation strategies are used, and if models are exclusively trained on simulated data and therefore never have seen real noise.[405] Furthermore, it is hoped that the attention that AI is attracting is not misused to "re-invent the wheel" by re-branding already established methods. In an optimistic scenario, however, AI and deep learning can be of great help in the automation of data analysis workflows, and the reduction of user-defined parameters. Examples include the deep learning-assisted component selection implemented in the PARADISe software, or the recently described deep learning-assisted versions of MCR.[254,255,406,407]

All in all, a lot of work remains to do, to transform analytical chemistry into an information science, with data analysis being just one part of this evolution. However, it has been shown

that chemometrics offers powerful tools that can help to bridge the gap between raw data acquisition and knowledge accumulation by extracting chemically meaningful information.

7 References

1. Cammann, K. *Competition-1st Prize Analytical Chemistry-Today's Definition and Interpretation*. *Fresenius J Anal Chem* vol. 343 (1992).
2. Perez-Bustamante, J. A. *Analytical Chemistry-Today's Definition and Interpretation*. *Journal of Fresenius J Anal Chem* vol. 343 (1992).
3. Zuckerman, A. M. *Analytical Chemistry-Today's Definition and Interpretation*. *Journal of Fresenius J Anal Chem* vol. 343 (1992).
4. Karayannis, M. I. & Efstathiou, C. E. Significant steps in the evolution of analytical chemistry - Is the today's analytical chemistry only chemistry? *Talanta* vol. **102**, 7–15 (2012).
5. Kowalski, B. R. Analytical chemistry as an information science. *TrAC Trends in Analytical Chemistry* vol. **1**, 71–74 (1981).
6. Wold, S., Sjostrom, M. *Chemometrics, Present and Future Success*. *Chemometrics and Intelligent Laboratory Systems* vol. 44 (1998).
7. Chanda, A. *et al.* Industry perspectives on process analytical technology: Tools and applications in API development. *Organic Process Research and Development* vol. 19 63–83 Preprint at https://doi.org/10.1021/op400358b (2015).
8. van den Berg, F., Lyndgaard, C. B., Sørensen, K. M. & Engelsen, S. B. Process Analytical Technology in the food industry. *Trends in Food Science and Technology* vol. 31 27–35 Preprint at https://doi.org/10.1016/j.tifs.2012.04.007 (2013).
9. Simon, L. L. *et al.* Assessment of recent process analytical technology (PAT) trends: A multiauthor review. *Organic Process Research and Development* vol. 19 3–62 Preprint at https://doi.org/10.1021/op500261y (2015).
10. Theodoridis, G. *et al.* Ensuring Fact-Based Metabolite Identification in Liquid Chromatography-Mass Spectrometry-Based Metabolomics. *Analytical Chemistry* vol. 95 3909–3916 Preprint at https://doi.org/10.1021/acs.analchem.2c05192 (2023).
11. Chaleckis, R., Meister, I., Zhang, P. & Wheelock, C. E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr Opin Biotechnol* vol. **55**, 44–50 (2019).
12. Alassali, A. *et al.* Towards higher quality of recycled plastics: Limitations from the material's perspective. *Sustainability (Switzerland)* vol. 13 Preprint at https://doi.org/10.3390/su132313266 (2021).
13. Santos, F. J. & Galceran, M. T. Modern developments in gas chromatography-mass spectrometry-based environmental analysis. *Journal of Chromatography A* vol. 1000 125–151 Preprint at https://doi.org/10.1016/S0021-9673(03)00305-4 (2003).
14. Tian, Z. *et al.* Suspect and Nontarget Screening for Contaminants of Emerging Concern in an Urban Estuary. *Environ Sci Technol* vol. **54**, 889–901 (2020).
15. Putri, S. P. *et al.* Application of gas chromatography-mass spectrometry-based metabolomics in food science and technology. *Journal of Bioscience and*

Bioengineering vol. 133 425–435 Preprint at https://doi.org/10.1016/j.jbiosc.2022.01.011 (2022).

16. Zoccali, M., Tranchida, P. Q. & Mondello, L. Fast gas chromatography-mass spectrometry: A review of the last decade. *TrAC Trends in Analytical Chemistry* vol. **118**, 444–452 (2019).

17. Farkas, T. & Chankvetadze, B. Ultrafast high-performance liquid chromatography. in *Liquid Chromatography* 145–176 (Elsevier, 2023). doi:10.1016/B978-0-323-99968-7.00031-X.

18. van den Hurk, R. S., Pursch, M., Stoll, D. R. & Pirok, B. W. J. Recent trends in two-dimensional liquid chromatography. *TrAC Trends in Analytical Chemistry* vol. **166**, 117166 (2023).

19. Muscalu, A. M. & Górecki, T. Comprehensive two-dimensional gas chromatography in environmental analysis. *TrAC - Trends in Analytical Chemistry* vol. 106 225–245 Preprint at https://doi.org/10.1016/j.trac.2018.07.001 (2018).

20. Stilo, F., Bicchi, C., Reichenbach, S. E. & Cordero, C. Comprehensive two-dimensional gas chromatography as a boosting technology in food-omic investigations. *Journal of Separation Science* vol. 44 1592–1611 Preprint at https://doi.org/10.1002/jssc.202100017 (2021).

21. Pollo, B. J., Alexandrino, G. L., Augusto, F. & Hantao, L. W. The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry. *TrAC Trends in Analytical Chemistry* vol. **105**, 202–217 (2018).

22. Yu, M., Yang, P., Song, H. & Guan, X. Research progress in comprehensive two-dimensional gas chromatography-mass spectrometry and its combination with olfactometry systems in the flavor analysis field. *Journal of Food Composition and Analysis* vol. 114 Preprint at https://doi.org/10.1016/j.jfca.2022.104790 (2022).

23. Parastar, H., Radović, J. R., Bayona, J. M. & Tauler, R. Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares ABC Highlights: Authored by Rising Stars and Top Experts. *Anal Bioanal Chem* vol. **405**, 6235–6249 (2013).

24. Skov, T. & Bro, R. Solving fundamental problems in chromatographic analysis. *Anal Bioanal Chem vol.* **390**, 281–285 (2008).

25. Gorrochategui, E., Jaumot, J. & Tauler, R. ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets. *BMC Bioinformatics* vol. **20**, 256 (2019).

26. Johnsen, L. G., Skou, P. B., Khakimov, B. & Bro, R. Gas chromatography – mass spectrometry data processing made easy. *J Chromatogr A* vol. **1503**, 57–64 (2017).

27. Olivieri, A. C., Neymeyr, K., Sawall, M. & Tauler, R. How noise affects the band boundaries in multivariate curve resolution. *Chemometrics and Intelligent Laboratory Systems* vol. **220**, 104472 (2022).

28. Olivieri, A. C. A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution − a tutorial. *Analytica Chimica Acta* vol. 1156 Preprint at https://doi.org/10.1016/j.aca.2021.338206 (2021).

29. Bortolato, S. A. & Olivieri, A. C. Chemometric processing of second-order liquid chromatographic data with UV-vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2. *Anal Chim Acta* vol. **842**, 11–19 (2014).

30. Ochoa, G. S., Sudol, P. E., Trinklein, T. J. & Synovec, R. E. Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. *Talanta* vol. **236**, (2022).

31. Yu, H. & Bro, R. PARAFAC2 and local minima. *Chemometrics and Intelligent Laboratory Systems* vol. **219**, (2021).

32. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* vol. **78**, 779–787 (2006).

33. Benton, H. P., Wong, D. M., Trauger, S. A. & Siuzdak, G. XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal Chem* vol. **80**, 6382–6389 (2008).

34. Tsugawa, H. *et al.* MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* vol. **12**, 523–526 (2015).

35. Heuckeroth, S. *et al.* Reproducible mass spectrometry data processing and compound annotation in MZmine 3. *Nat Protoc* (2024) doi:10.1038/s41596-024-00996-y.

36. Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology* vol. 41 447–449 Preprint at https://doi.org/10.1038/s41587-023-01690-2 (2023).

37. Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nat Biotechnol* vol. **38**, 1159–1163 (2020).

38. Guijas, C. *et al.* METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal Chem* vol. **90**, 3156–3164 (2018).

39. Gürdeniz, G., Kristensen, M., Skov, T. & Dragsted, L. O. The effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats. *Metabolites* vol. **2**, 77–99 (2012).

40. Hohrenk, L. L. *et al.* Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Anal Chem* vol. **92**, 1898–1907 (2020).

41. Wang, X. C. *et al.* A comparison of feature extraction capabilities of advanced UHPLC-HRMS data analysis tools in plant metabolomics. *Anal Chim Acta* vol. **1254**, (2023).

42. Rafiei, A. & Sleno, L. Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Communications in Mass Spectrometry* vol. **29**, 119–127 (2015).

43. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal Chem* vol. **89**, 8689–8695 (2017).

44. Stefanuto, P. H., Smolinska, A. & Focant, J. F. Advanced chemometric and data handling tools for GC×GC-TOF-MS: Application of chemometrics and related advanced data handling in chemical separations. *TrAC - Trends in Analytical Chemistry* vol. 139 Preprint at https://doi.org/10.1016/j.trac.2021.116251 (2021).

45. Trinklein, T. J. *et al.* Recent Advances in GC×GC and Chemometrics to Address Emerging Challenges in Nontargeted Analysis. *Anal Chem* vol. **95**, 264–286 (2023).

46. Pirok, B. W. J., Stoll, D. R. & Schoenmakers, P. J. Recent Developments in Two-Dimensional Liquid Chromatography: Fundamental Improvements for Practical Applications. *Analytical Chemistry* vol. 91 240–263 Preprint at https://doi.org/10.1021/acs.analchem.8b04841 (2019).

47. Gallagher, N. Documentation ALS-SIT in PLS-Toolbox, Eigenvector Inc. https://wiki.eigenvector.com/index.php?title=Als_sit, accessed 31.08.2024 (2023).

48. Quintanilla Casas, B., Bro, R., Hinrich, J. & Davie-Martin, C. Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identification System for processing GC–MS data. (2023) doi:https://doi.org/10.21203/rs.3.pex-2143/v1.

49. Solle, D. Be FAIR to your data. *Anal Bioanal Chem* vol. **412**, 3961–3965 (2020).

50. Juchli, D. SiLA 2: The Next Generation Lab Automation Standard. *Adv Biochem Eng Biotechnol* vol. 182 pp. 147–174 (2022). doi:10.1007/10_2022_204.

51. Kayser, H. & Lau, M.-L. Growing value of data standardization: Allotrope Foundation Connect Workshop Proceedings. *Drug Discov Today* vol. **29**, 103988 (2024).

52. Oliver, S. Systematic functional analysis of the yeast genome. *Trends Biotechnol* vol. **16**, 373–378 (1998).

53. Fiehn, O. *et al. Metabolite Profiling for Plant Functional Genomics*. *Nature Biotechnology*, vol. 18, 1157–1161 http://biotech.nature.com (2000).

54. Cajka, T. & Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Analytical Chemistry* vol. 88 524–545 Preprint at https://doi.org/10.1021/acs.analchem.5b04491 (2016).

55. Alseekh, S. & Fernie, A. R. Metabolomics 20 years on: what have we learned and what hurdles remain? *Plant Journal* vol. 94, 933–942 Preprint at https://doi.org/10.1111/tpj.13950 (2018).

56. Poulsen, R. *et al.* A case study of PAH contamination using blue mussels as a bioindicator in a small Greenlandic fishing harbor. *Mar Pollut Bull* vol. **171**, 112688 (2021).

57. Kronik, O. M. *et al.* A study of the spatial distribution patterns of airborne polycyclic aromatic hydrocarbons in crowberry (Empetrum nigrum) in Ilulissat, Greenland. Environ Sci Pollut Res vol. 28, 23133–23142 doi:10.1007/s11356-021-12365-3/Published.

58. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *Analytical Procedure Development (Q14)*. https://database.ich.org/sites/default/files/ICH_Q14_Document_Step2_Guideline_2022_0324.pdf, accessed 31.08.2024 (2022).

59. Food and Drug Administration (FDA). *Analytical Procedures and Methods Validation for Drugs and Biologics*. https://www.fda.gov/media/87801/download , accessed 31.08.2024 (2015).

60. Magnusson, B. Ö. U. Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics. https://www.eurachem.org/images/stories/Guides/pdf/MV_guide_2nd_ed_EN.pdf , accessed 31.08.2024 (2014).

61. Jimidar, M. I., Heylen, P. & De Smet, M. 16 Method validation. *Separation Science and Technology* vol. 8, 441–458 (2007). doi:10.1016/S0149-6395(07)80022-5.

62. Thompson, M., Ellison, S. L. R. & Wood, R. Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC Technical Report). *Pure and Applied Chemistry* **74**, 835–855 (2002).

63. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted metabolomics. *Curr Protoc Mol Biol* (2012). doi: 10.1002/0471142727.mb3002s98

64. Kirwan, J. A. *et al.* Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. *Metabolomics* vol. **18**, 70 (2022).

65. Broadhurst, D. *et al.* Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* vol. 14 Preprint at https://doi.org/10.1007/s11306-018-1367-3 (2018).

66. Regalado, E. L., Dermenjian, R. K., Joyce, L. A. & Welch, C. J. Detection of dehalogenation impurities in organohalogenated pharmaceuticals by UHPLC–DAD–HRESIMS. *J Pharm Biomed Anal* **92**, 1–5 (2014).

67. Giera, M., Yanes, O. & Siuzdak, G. Metabolite discovery: Biochemistry's scientific driver. *Cell Metabolism* vol. 34 21–34 Preprint at https://doi.org/10.1016/j.cmet.2021.11.005 (2022).

68. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J Am Soc Mass Spectrom* vol. **27**, 1897–1905 (2016).

69. Hollender, J., Schymanski, E. L., Singer, H. P. & Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ Sci Technol* vol. **51**, 11505–11512 (2017).

70. Gertsman, I. & Barshop, B. A. Promises and pitfalls of untargeted metabolomics. *Journal of Inherited Metabolic Disease* vol. 41 355–366 Preprint at https://doi.org/10.1007/s10545-017-0130-7 (2018).

71. Hollender, J. *et al.* NORMAN guidance on suspect and non-target screening in environmental monitoring. *Environmental Sciences Europe* vol. 35 Preprint at https://doi.org/10.1186/s12302-023-00779-4 (2023).

72. Cain, C. N., Trinklein, T. J., Ochoa, G. S. & Synovec, R. E. Tile-Based Pairwise Analysis of GC × GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification. *Anal Chem* vol. **94**, 5658–5666 (2022).

73. Cain, C. N., Sudol, P. E., Berrier, K. L. & Synovec, R. E. Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis. *Talanta* vol. **233**, (2021).

74. Jungclaus, G., Avila, V. & Hites, R. Organic compounds in an industrial Wastewater: a case study of their environmental impact. *Environ Sci Technol* vol. **12**, 88–96 (1978).

75. Sheldon, L. S. & Hites, R. A. Organic compounds in the Delaware River. *Environ Sci Technol* vol. **12**, 1188–1194 (1978).

76. Tisler, S., Kilpinen, K., Pattison, D. I., Tomasi, G. & Christensen, J. H. Quantitative Nontarget Analysis of CECs in Environmental Samples Can Be Improved by Considering All Mass Adducts. *Anal Chem* vol. **96**, 229–237 (2024).

77. Wang, Y.-Q. *et al.* Per- and polyfluoralkyl substances (PFAS) in drinking water system: Target and non-target screening and removal assessment. *Environ Int* vol. **163**, 107219 (2022).

78. Yore, M. M. *et al.* Discovery of a Class of Endogenous Mammalian Lipids with Anti-Diabetic and Anti-inflammatory Effects. *Cell* vol. **159**, 318–332 (2014).

79. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* vol. **472**, 57–63 (2011).

80. Shewry, P. R. *et al.* Are GM and conventionally bred cereals really different? *Trends Food Sci Technol* vol. **18**, 201–209 (2007).

81. Samuelsson, L. M. & Larsson, D. G. J. Contributions from metabolomics to fish research. *Mol Biosyst* vol. **4**, 974 (2008).

82. Vosough, M., Schmidt, T. C. & Renner, G. Non-target screening in water analysis: recent trends of data evaluation, quality assurance, and their future perspectives. *Anal Bioanal Chem* vol. **416**, 2125–2136 (2024).

83. Schwarzenbach, R. P. *et al.* The Challenge of Micropollutants in Aquatic Systems. *Science (1979)* vol. **313**, 1072–1077 (2006).

84. Xia, J. & Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc* vol. **6**, 743–760 (2011).

85. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: A web-based platform to process untargeted metabolomic data. *Anal Chem* vol. **84**, 5035–5039 (2012).

86. Lommen, A. & Kools, H. J. MetAlign 3.0: Performance enhancement by efficient use of advances in computer hardware. *Metabolomics* vol. **8**, 719–726 (2012).

87. Helmus, R., ter Laak, T. L., van Wezel, A. P., de Voogt, P. & Schymanski, E. L. patRoon: open source software platform for environmental mass spectrometry based non-target screening. *J Cheminform* vol. **13**, (2021).

88. Beatriz Quintanilla-Casas, R. B. J. L. H. et al. *Tutorial on PARADISe: PARAFAC2-Based Deconvolution and Identification System for Processing GC–MS Data*. (2023).

89. Mikaliunaite, L. & Synovec, R. E. Computational method for untargeted determination of cycling yeast metabolites using comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *Talanta* vol. **244**, (2022).

90. Pollo, B. J. *et al.* Chemometrics, Comprehensive Two-Dimensional gas chromatography and “omics” sciences: Basic tools and recent applications. *TrAC - Trends in Analytical Chemistry* vol. 134 Preprint at https://doi.org/10.1016/j.trac.2020.116111 (2021).

91. Ulrich, E. M. *et al.* EPA’s non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. *Anal Bioanal Chem* vol. **411**, 853–866 (2019).

92. Schymanski, E. L. *et al.* Non-target screening with high-resolution mass spectrometry: Critical review using a collaborative trial on water analysis. *Anal Bioanal Chem* vol. **407**, 6237–6255 (2015).

93. Rostkowski, P. *et al.* The strength in numbers: comprehensive characterization of house dust using complementary mass spectrometric techniques. *Anal Bioanal Chem* vol. **411**, 1957–1977 (2019).

94. Clark, T. N. *et al.* Interlaboratory Comparison of Untargeted Mass Spectrometry Data Uncovers Underlying Causes for Variability. *J Nat Prod* vol. **84**, 824–835 (2021).

95. Lin, Y., Caldwell, G. W., Li, Y., Lang, W. & Masucci, J. Inter-laboratory reproducibility of an untargeted metabolomics GC–MS assay for analysis of human plasma. *Sci Rep* vol. **10**, 10918 (2020).

96. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* vol. **3**, 211–221 (2007).

97. Renner, G. & Reuschenbach, M. Critical review on data processing algorithms in non-target screening: challenges and opportunities to improve result comparability. *Analytical and Bioanalytical Chemistry* vol. 415 4111–4123 Preprint at https://doi.org/10.1007/s00216-023-04776-7 (2023).

98. Considine, E. C., Thomas, G., Boulesteix, A. L., Khashan, A. S. & Kenny, L. C. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics* vol. **14**, 7 (2018).

99. Place, B. J. *et al.* An Introduction to the Benchmarking and Publications for Non-Targeted Analysis Working Group. *Anal Chem* vol. **93**, 16289–16296 (2021).

100. Du, B. *et al.* Development of suspect and non-target screening methods for detection of organic contaminants in highway runoff and fish tissue with high-resolution time-of-flight mass spectrometry. *Environ Sci Process Impacts* vol. **19**, 1185–1196 (2017).

101. Manz, K. E. *et al.* Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome. *Journal of Exposure Science and Environmental Epidemiology* vol. 33 524–536 Preprint at https://doi.org/10.1038/s41370-023-00574-6 (2023).

102. Gallidabino, M. D., Hamdan, L., Murphy, B. & Barron, L. P. Suspect screening of halogenated carboxylic acids in drinking water using ion exchange chromatography – High resolution (Orbitrap) mass spectrometry (IC-HRMS). *Talanta* vol. **178**, 57–68 (2018).

103. Malm, L. *et al.* Guide to semi-quantitative non-targeted screening using lc/esi/hrms. *Molecules* vol. **26**, (2021).

104. Tisler, S. *et al.* From data to reliable conclusions: Identification and comparison of persistent micropollutants and transformation products in 37 wastewater samples by non-target screening prioritization. *Water Res* vol. **219**, (2022).

105. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat Biotechnol* vol. **38**, 23–26 (2020).

106. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* vol. **58**, 109–130 (2001).

107. Mayerhöfer, T. G., Pahlow, S. & Popp, J. The Bouguer-Beer-Lambert Law: Shining Light on the Obscure. *ChemPhysChem* vol. **21**, 2029–2046 (2020).

108. Ettre, L. S. Nomenclature for chromatography (IUPAC Recommendations 1993). *Pure and Applied Chemistry* vol. **65**, 819–872 (1993).

109. Kaiser, J. & Chalk, S. *The IUPAC Compendium of Chemical Terminology*. (International Union of Pure and Applied Chemistry (IUPAC), Research Triangle Park, NC, 2024). doi:10.1351/goldbook.

110. Miller, J. Introduction to Chromatography. in *Chromatography: Concepts and Contrasts* vol. 2, pp 35–66 (Wiley, 2009). doi:10.1002/9780471980582.ch2.

111. Komsta, L., Waksmundzka-Hajnos, M., Sherma, J. *Thin Layer Chromatography in Drug Analysis*. vol. 1 (CRC Press, 2013). doi: https://doi.org/10.1201/b15637.

112. Ettre, L. S. Chromatography: The separation technique of the 20th century. *Chromatographia* vol. **51**, 7–17 (2000).

113. James, A. T. & Martin, A. J. P. Gas-liquid partition chromatography: the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid. *Biochemical Journal* vol. **50**, 679–690 (1952).

114. James, A. T., Martin, A. J. P. & Smith, G. H. Gas-liquid partition chromatography: the separation and micro-estimation of ammonia and the methylamines. *Biochemical Journal* vol. **52**, 238–242 (1952).

115. Martin, A. J. P. & Synge, R. L. M. Separation of the higher monoamino-acids by counter-current liquid-liquid extraction: the amino-acid composition of wool. *Biochemical Journal* vol. **35**, 91–121 (1941).

116. Martin, A. J. P. & Synge, R. L. M. A new form of chromatogram employing two liquid phases. *Biochemical Journal* vol. **35**, 1358–1368 (1941).

117. Horvath, C. G., Lipsky, R. S. Use of Liquid Ion Exchange Chromatography for the Separation of Organic Compounds. *Nature* vol. **211**, 748–749 (1966).

118. Horvath, C. G., Preiss, B. A. & Lipsky, S. R. Fast liquid chromatography. Investigation of operating parameters and the separation of nucleotides on pellicular ion exchangers. *Anal Chem* vol. **39**, 1422–1428 (1967).

119. Dunn, W. B. & Ellis, D. I. Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry* vol. **24**, 285–294 (2005).

120. Bushey, M. M. & Jorgenson, J. W. Automated instrumentation for comprehensive two-dimensional high-performance liquid chromatography of proteins. *Anal Chem* vol. **62**, 161–167 (1990).

121. Liu, Z. & Phillips, J. B. Comprehensive Two-Dimensional Gas Chromatography using an On-Column Thermal Modulator Interface. *J Chromatogr Sci* vol. **29**, 227–231 (1991).

122. Giddings, J. C. Two-dimensional separations: concept and promise. *Anal Chem* vol. **56**, 1258A-1270A (1984).

123. Giddings, J. C. Concepts and comparisons in multidimensional separation. *Journal of High Resolution Chromatography* vol. **10**, 319–323 (1987).

124. Gruber, B. *et al.* Comprehensive two-dimensional gas chromatography in forensic science: A critical review of recent trends. *TrAC Trends in Analytical Chemistry* vol. **105**, 292–301 (2018).

125. Aspromonte, J., Wolfs, K. & Adams, E. Current application and potential use of GC × GC in the pharmaceutical and biomedical field. *J Pharm Biomed Anal* vol. **176**, 112817 (2019).

126. Desfontaine, V., Guillarme, D., Francotte, E. & Nováková, L. Supercritical fluid chromatography in pharmaceutical analysis. *J Pharm Biomed Anal* vol. **113**, 56–71 (2015).

127. Taylor, L. T. Supercritical fluid chromatography for the 21st century. *J Supercrit Fluids* vol. **47**, 566–573 (2009).

128. Dorman, F. L. & Dawes, P. Column technology. in *Gas Chromatography* 99–116 (Elsevier, 2021). doi:10.1016/B978-0-12-820675-1.00003-4.

129. Poole, C. F. Reversed-phase liquid chromatography. in *Liquid Chromatography* 89–119 (Elsevier, 2023). doi:10.1016/B978-0-323-99968-7.00015-1.

130. Habgood, H. W. & Harris, W. E. Retention Temperature and Column Efficiency in Programmed Temperature Gas Chromatography. *Anal Chem* vol. **32**, 450–453 (1960).

131. Merrick, M. & Blumberg, L. M. Optimal heating rate in constant pressure and constant flow gas chromatography. *J Sep Sci* vol. **44**, 3254–3267 (2021).

132. García-Álvarez-Coque, M. C., Torres-Lapasió, J. R. & Baeza-Baeza, J. J. Models and objective functions for the optimisation of selectivity in reversed-phase liquid chromatography. *Anal Chim Acta* vol. **579**, 125–145 (2006).

133. Nikitas, P. & Pappa-Louisi, A. Retention models for isocratic and gradient elution in reversed-phase liquid chromatography. *J Chromatogr A* vol. **1216**, 1737–1755 (2009).

134. Blumberg, L. M. Theory of gas chromatography. in *Gas Chromatography* 19–97 (Elsevier, 2021). doi:10.1016/B978-0-12-820675-1.00026-5.

135. Davis, J. M. & Giddings, J. Calvin. Statistical theory of component overlap in multicomponent chromatograms. *Anal Chem* vol. **55**, 418–424 (1983).

136. Felinger, A. & Cavazzini, A. Kinetic theories of liquid chromatography. in *Liquid Chromatography* 17–37 (Elsevier, 2017). doi:10.1016/B978-0-12-805393-5.00002-6.

137. Blumberg, L. M. Theory of gas chromatography. in *Gas Chromatography* 19–97 (Elsevier, 2021). doi:10.1016/B978-0-12-820675-1.00026-5.

138. Neue, U. D. & Kuss, H.-J. Improved reversed-phase gradient retention modeling. *J Chromatogr A* **1217**, 3794–3803 (2010).

139. den Uijl, M. J., Schoenmakers, P. J., Pirok, B. W. J. & van Bommel, M. R. Recent applications of retention modelling in liquid chromatography. *J Sep Sci* vol. **44**, 88–114 (2021).

140. Wiczling, P. & Kaliszan, R. Influence of ph on retention in linear organic modifier gradient RP HPLC. *Anal Chem* vol. **80**, 7855–7861 (2008).

141. Dolan, J. W. *Temperature Selectivity in Reversed-Phase High Performance Liquid Chromatography*. *Journal of Chromatography A* www.elsevier.com/locate/chroma (2002).

142. Macko, T. & Berek, D. Pressure effects in HPLC: Influence of pressure and pressure changes on peak shape, base line, and retention volume in HPLC separations. *J Liq Chromatogr Relat Technol* vol. **24**, 1275–1293 (2001).

143. Dolan, J. W. & Snyder, L. R. Theory and practice of gradient elution liquid chromatography. in *Liquid Chromatography* 389–402 (Elsevier, 2017). doi:10.1016/B978-0-12-805393-5.00015-4.

144. van Deemter, J. J., Zuiderweg, F. J. & Klinkenberg, A. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. *Chem Eng Sci* vol. **5**, 271–289 (1956).

145. Blumberg, L. M. Extension of Golay plate height equation for open-tubular columns. *J Chromatogr A* vol. **1524**, 303–306 (2017).

146. Fuller, E. N., Schettler, P. D. & Giddings, J. Calvin. NEW METHOD FOR PREDICTION OF BINARY GAS-PHASE DIFFUSION COEFFICIENTS. *Ind Eng Chem* vol. **58**, 18–27 (1966).

147. Horváth, S., Lukács, D., Farsang, E. & Horváth, K. Unbiased Determination of Adsorption Isotherms by Inverse Method in Liquid Chromatography. *Molecules* vol. **28**, 1031 (2023).

148. García-Álvarez-Coque, M. C., Torres-Lapasió, J. R. & Navarro-Huerta, J. A. Secondary chemical equilibria in reversed-phase liquid chromatography. in *Liquid Chromatography* 125–146 (Elsevier, 2017). doi:10.1016/B978-0-12-805393-5.00005-1.

149. Nawrocki, J. The silanol group and its role in liquid chromatography. *J Chromatogr A* vol. **779**, 29–71 (1997).

150. Kruve, A. & Kaupmees, K. Adduct Formation in ESI/MS by Mobile Phase Additives. *J Am Soc Mass Spectrom* vol. **28**, 887–894 (2017).

151. Taylor, P. J. Matrix effects: The Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry. *Clinical Biochemistry* vol. 38 328–334 Preprint at https://doi.org/10.1016/j.clinbiochem.2004.11.007 (2005).

152. Broyles, B. S., Shalliker, R. A., Cherrak, D. E. & Guiochon, G. *Visualization of Viscous Fingering in Chromatographic Columns*. *Journal of Chromatography A* vol. 822 (1998).

153. Keunchkarian, S., Reta, M., Romero, L. & Castells, C. Effect of sample solvent on the chromatographic peak shape of analytes eluted under reversed-phase liquid chromatogaphic conditions. *J Chromatogr A* vol. **1119**, 20–28 (2006).

154. van den Hurk, R. S., Pursch, M., Stoll, D. R. & Pirok, B. W. J. Recent trends in two-dimensional liquid chromatography. *TrAC - Trends in Analytical Chemistry* vol. 166 Preprint at https://doi.org/10.1016/j.trac.2023.117166 (2023).

155. Chen, Y., Li, J. & Schmitz, O. J. Development of an At-Column Dilution Modulator for Flexible and Precise Control of Dilution Factors to Overcome Mobile Phase Incompatibility in Comprehensive Two-Dimensional Liquid Chromatography. *Anal Chem* vol. **91**, 10251–10257 (2019).

156. Stoll, D. R., Shoykhet, K., Petersson, P. & Buckenmaier, S. Active Solvent Modulation: A Valve-Based Approach To Improve Separation Compatibility in Two-Dimensional Liquid Chromatography. *Anal Chem* vol. **89**, 9260–9267 (2017).

157. Tang, S. & Venkatramani, C. J. Resolving Solvent Incompatibility in Two-Dimensional Liquid Chromatography with In-Line Mixing Modulation. *Anal Chem* vol. **94**, 16142–16150 (2022).

158. Carr, P. W. & Stoll, D. R. A Study of Peak Capacity Optimization in One-Dimensional Gradient Elution Reversed-Phase Chromatography. in *Advances in Chromatography* 1–21 (CRC Press, 2017). doi:10.1201/9781315158075-1.

159. Stoll, D. R. *et al.* Fast, comprehensive two-dimensional liquid chromatography. *J Chromatogr A* vol. **1168**, 3–43 (2007).

160. Prebihalo, S. E. *et al.* Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications. *Analytical Chemistry* vol. 90 505–532 Preprint at https://doi.org/10.1021/acs.analchem.7b04226 (2018).

161. Parastar, H. & Tauler, R. Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: A new insight to address current chromatographic challenges. *Anal Chem* vol. **86**, 286–297 (2014).

162. Pérez-Cova, M., Jaumot, J. & Tauler, R. Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches. *TrAC - Trends in Analytical Chemistry* vol. 137 Preprint at https://doi.org/10.1016/j.trac.2021.116207 (2021).

163. Navarro-Reig, M., Jaumot, J. & Tauler, R. An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis. *J Chromatogr A* vol. **1568**, 80–90 (2018).

164. Gohlke, R. S. & McLafferty, F. W. Early gas chromatography/mass spectrometry. *J Am Soc Mass Spectrom* vol. **4**, 367–371 (1993).

165. Bleakney, W. A New Method of Positive Ray Analysis and Its Application to the Measurement of Ionization Potentials in Mercury Vapor. *Physical Review* vol. **34**, 157–160 (1929).

166. Nier, A. O. A Mass Spectrometer for Isotope and Gas Analysis. *Review of Scientific Instruments* vol. **18**, 398–411 (1947).

167. Stephens, W. A Pulsed Mass Spectrometer with Time Dispersion. *Physical Review* vol. **69**, 674–674 (1946).

168. Gomer, R. & Inghram, M. G. Applications of Field Ionization to Mass Spectrometry. *J Am Chem Soc* vol. **77**, 500–500 (1955).

169. Munson, M. S. B. & Field, F. H. Chemical Ionization Mass Spectrometry. I. General Introduction. *J Am Chem Soc* vol. **88**, 2621–2630 (1966).

170. Vinaixa, M. *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC - Trends in Analytical Chemistry* vol. 78 23–35 Preprint at https://doi.org/10.1016/j.trac.2015.09.005 (2016).

171. Arpino, P. & Paristech, C. *History of LC-MS Development and Interfacing LC-MS: Principles and Instrumentation History of LC-MS Development and Interfacing*. https://www.researchgate.net/publication/291339319.

172. Horning, E. C., Horning, M. G., Carroll, D. I., Dzidic, I. & Stillwell, R. N. New picogram detection system based on a mass spectrometer with an external ionization source at atmospheric pressure. *Anal Chem* vol. **45**, 936–943 (1973).

173. Carroll, D. I., Dzidic, I., Stillwell, R. N., Haegele, K. D. & Horning, E. C. Atmospheric pressure ionization mass spectrometry. Corona discharge ion source for use in a liquid chromatograph-mass spectrometer-computer analytical system. *Anal Chem* vol. **47**, 2369–2373 (1975).

174. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. *Electrospray Ionization for Mass Spectrometry of Large Biomolecules*. *New Series* vol. 246 (1989).

175. Dole, M. *et al.* Molecular Beams of Macroions. *J Chem Phys* vol. **49**, 2240–2249 (1968).

176. Gross, J. H. Electrospray Ionization. in *Mass Spectrometry* 721–778 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-54398-7_12.

177. Paul, W. & Steinwedel, H. Notizen: Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift für Naturforschung A* vol. **8**, 448–450 (1953).

178. Paul, W. & Raether, M. Das elektrische Massenfilter. *Zeitschrift fuer Physik* vol. **140**, 262–273 (1955).

179. Smith, L. G. A New Magnetic Period Mass Spectrometer. *Review of Scientific Instruments* vol. **22**, 115–116 (1951).

180. Sommer, H., Thomas, H. A. & Hipple, J. A. The Measurement of e/M by Cyclotron Resonance. *Physical Review* vol. **82**, 697–702 (1951).

181. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* vol. **40**, 430–443 (2005).

182. Hoang, C. *et al.* Tandem Mass Spectrometry across Platforms. *Anal Chem* vol. **96**, 5478–5488 (2024).

183. Gross, J. H. Instrumentation. in *Mass Spectrometry* 151–292 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-54398-7_4.

184. Cech, N. B. & Enke, C. G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom Rev* vol. **20**, 362–387 (2001).

185. Gross, J. H. Chemical Ionization. in *Mass Spectrometry* 439–496 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-54398-7_7.

186. Gross, J. H. Practical Aspects of Electron Ionization. in *Mass Spectrometry* 293–324 (Springer International Publishing, 2017). doi:10.1007/978-3-319-54398-7_5.

187. Capellades, J. *et al.* Exploring the use of gas chromatography coupled to chemical ionization mass spectrometry (GC-CI-MS) for stable isotope labeling in metabolomics. *Anal Chem* vol. **93**, 1242–1248 (2021).

188. Xian, F., Hendrickson, C. L. & Marshall, A. G. High Resolution Mass Spectrometry. *Anal Chem* vol. **84**, 708–719 (2012).

189. Hawkridge, A. M. Practical Considerations and Current Limitations in Quantitative Mass Spectrometry-based Proteomics. in *Quantitative Proteomics* 1–25 (The Royal Society of Chemistry, 2014). doi:10.1039/9781782626985-00001.

190. Chernushevich, I. V., Merenbloom, S. I., Liu, S. & Bloomfield, N. A W-Geometry Ortho-TOF MS with High Resolution and Up to 100% Duty Cycle for MS/MS. *J Am Soc Mass Spectrom* vol. **28**, 2143–2150 (2017).

191. Willis, P., Jaloszynski, J. & Artaev, V. Improving duty cycle in the Folded Flight Path high-resolution time-of-flight mass spectrometer. *Int J Mass Spectrom* vol. **459**, (2021).

192. Harvey, D. J. Mass spectrometric detectors for gas chromatography. in *Gas Chromatography* 399–424 (Elsevier, 2021). doi:10.1016/B978-0-12-820675-1.00022-8.

193. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. *Anal Chem* vol. **92**, 8072–8080 (2020).

194. Fenaille, F., Barbier Saint-Hilaire, P., Rousseau, K. & Junot, C. Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: Where do we stand? *Journal of Chromatography A* vol. 1526 1–12 Preprint at https://doi.org/10.1016/j.chroma.2017.10.043 (2017).

195. Plumb, R. S. *et al.* UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry* vol. **20**, 1989–1994 (2006).

196. Doerr, A. DIA mass spectrometry. *Nature Methods* vol. 12 35 Preprint at https://doi.org/10.1038/nmeth.3234 (2014).

197. Kitteringham, N. R., Jenkins, R. E., Lane, C. S., Elliott, V. L. & Park, B. K. Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* vol. 877 1229–1239 Preprint at https://doi.org/10.1016/j.jchromb.2008.11.013 (2009).

198. Kronik, O. M., Liang, X., Nielsen, N. J., Christensen, J. H. & Tomasi, G. Obtaining clean and informative mass spectra from complex chromatographic and high-resolution all-ions-fragmentation data by nonnegative parallel factor analysis 2. *J Chromatogr A* vol. **1682**, 463501 (2022).

199. Malla, M. A. *et al.* GC–MS based untargeted metabolomics reveals the metabolic response of earthworm (Eudrilus eugeniae) after chronic combinatorial exposure to three different pesticides. *Sci Rep* vol. **13**, (2023).

200. Wong, J. W. *et al.* Multiresidue pesticide analysis in fresh produce by capillary gas chromatography-mass spectrometry/selective ion monitoring (GC-MS/SIM) and -tandem mass spectrometry (GC-MS/MS). *J Agric Food Chem* vol. **58**, 5868–5883 (2010).

201. Samokhin, A. Spectral skewing in gas chromatography–mass spectrometry: Misconceptions and realities. *J Chromatogr A* vol. **1576**, 113–119 (2018).

202. Schauer, N. *et al.* GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett* vol. **579**, 1332–1337 (2005).

203. Purcaro, G. *et al.* Evaluation of a rapid-scanning quadrupole mass spectrometer in an apolar × ionic-liquid comprehensive two-dimensional gas chromatography system. *Anal Chem* vol. **82**, 8583–8590 (2010).

204. Adahchour, M. *et al.* Comprehensive two-dimensional gas chromatography coupled to a rapid-scanning quadrupole mass spectrometer: Principles and applications. *J Chromatogr A* vol. **1067**, 245–254 (2005).

205. Kirchner, M., Matisová, E., Hrouzková, S. & De Zeeuw, J. Possibilities and limitations of quadrupole mass spectrometric detector in fast gas chromatography. *J Chromatogr A* vol. **1090**, 126–132 (2005).

206. Zhang, P., Carlin, S., Franceschi, P., Mattivi, F. & Vrhovsek, U. Application of a Target-Guided Data Processing Approach in Saturated Peak Correction of GC×GC Analysis. *Anal Chem* vol. **94**, 1941–1948 (2022).

207. Wei, A. A. J., Joshi, A., Chen, Y. & McIndoe, J. S. Strategies for avoiding saturation effects in ESI-MS. *Int J Mass Spectrom* vol. **450**, (2020).

208. Tang, K., Page, J. S. & Smith, R. D. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J Am Soc Mass Spectrom* vol. **15**, 1416–1423 (2004).

209. Ladak, A., Amit Gujar, ;, Anderson, T., Kenneth, ; & Free, M. Extended Linear Dynamic Range with a New Electron Multiplier System on Single Quadrupole GC-MS. *Thermo Fisher* https://assets.thermofisher.cn/TFS-Assets/CMD/Technical-Notes/tn-000421-gc-ms-linear-dynamic-range-xlxr-detector-tn000421-en.pdf, accessed 31.08.2024 (2022).

210. Bilbao, A. *et al.* An algorithm to correct saturated mass spectrometry ion abundances for enhanced quantitation and mass accuracy in omic studies. *Int J Mass Spectrom* vol. **427**, 91–99 (2018).

211. Chernushevich, I. V., Loboda, A. V. & Thomson, B. A. An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry* vol. **36**, 849–865 (2001).

212. Moulds, R., Worthington, K., Doorbar, P., Kenny, D.J., Pringle, S., Morphet, J., Douce, D.. Extending the Linear Range of Quadrupole Detectors. *Waters Cooperation* https://www.waters.com/webassets/cms/library/docs/asms2014_moulds_quadrupole_detectors.pdf accessed 31.08.2024 (2014).

213. Yuan, L., Zhang, D., Jemal, M. & Aubry, A. Systematic evaluation of the root cause of non-linearity in liquid chromatography/tandem mass spectrometry bioanalytical assays and strategy to predict and extend the linear standard curve range. *Rapid Communications in Mass Spectrometry* vol. **26**, 1465–1474 (2012).

214. Alfaro, C. M., Uwakweh, A.-O., Todd, D. A., Ehrmann, B. M. & Cech, N. B. Investigations of Analyte-Specific Response Saturation and Dynamic Range Limitations in Atmospheric Pressure Ionization Mass Spectrometry. *Anal Chem* vol. **86**, 10639–10645 (2014).

215. Major, H. J., Castro-Perez, J. M., Hoyes, J. & Harland, G. Extending the Quantitative Linear Range of a Quadrupole-Time of Flight Mass Spectrometer Using Digital Deadtime Correction (DDTC). *Waters Cooperation* https://www.waters.com/content/dam/waters/en/app-notes/2001/AB39/AB39-en.pdf, accessed 31.08.2024 (2001).

216. Sorochan Armstrong, M. D., Hinrich, J. L., de la Mata, A. P. & Harynuk, J. J. PARAFAC2×N: Coupled decomposition of multi-modal data with drift in N modes. *Anal Chim Acta* vol. **1249**, (2023).

217. Weggler, B. A. *et al.* A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software. *J Chromatogr A* vol. **1635**, (2021).

218. Hunter, C. How can I tell if my Spectral Peak is saturated on the TripleTOF system? *Sciex* https://community.sciex.com/2021/08/08/how-can-i-tell-if-my-spectral-peak-is-saturated-on-the-tripletof-system/, accessed 31.08.2024 (2021).

219. Annesley, T. M. Ion Suppression in Mass Spectrometry. *Clin Chem* vol. **49**, 1041–1044 (2003).

220. Bonfiglio, R., King, R. C., Olah, T. V. & Merkle, K. The effects of sample preparation methods on the variability of the electrospray ionization response for model drug compounds. *Rapid Communications in Mass Spectrometry* vol. **13**, 1175–1185 (1999).

221. Furey, A., Moriarty, M., Bane, V., Kinsella, B. & Lehane, M. Ion suppression; A critical review on causes, evaluation, prevention and applications. *Talanta* vol. 115 104–122 Preprint at https://doi.org/10.1016/j.talanta.2013.03.048 (2013).

222. Matuszewski, B. K., Constanzer, M. L. & Chavez-Eng, C. M. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Anal Chem* vol. **75**, 3019–3030 (2003).

223. Tisler, S., Pattison, D. I. & Christensen, J. H. Correction of Matrix Effects for Reliable Non-target Screening LC–ESI–MS Analysis of Wastewater. *Anal Chem* vol. **93**, 8432–8441 (2021).

224. Vereyken, L., Dillen, L., Vreeken, R. J. & Cuyckens, F. High-Resolution Mass Spectrometry Quantification: Impact of Differences in Data Processing of Centroid and Continuum Data. *J Am Soc Mass Spectrom* vol. **30**, 203–212 (2019).

225. Samanipour, S. *et al.* From Centroided to Profile Mode: Machine Learning for Prediction of Peak Width in HRMS Data. *Anal Chem* vol. **93**, 16562–16570 (2021).

226. Gorrochategui, E., Jaumot, J., Lacorte, S. & Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC - Trends in Analytical Chemistry* vol. 82 425–442 Preprint at https://doi.org/10.1016/j.trac.2016.07.004 (2016).

227. Urban, J., Afseth, N. K. & Štys, D. Fundamental definitions and confusions in mass spectrometry about mass assignment, centroiding and resolution. *TrAC - Trends in Analytical Chemistry* vol. 53 126–136 Preprint at https://doi.org/10.1016/j.trac.2013.07.010 (2014).

228. Wang, Y. & Gu, M. The Concept of Spectral Accuracy for MS. *Anal Chem* vol. **82**, 7055–7062 (2010).

229. Du, X., Smirnov, A., Pluskal, T., Jia, W. & Sumner, S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. in *Methods in Molecular Biology* vol. 2104 25–48 (Humana Press Inc., 2020).

230. Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* vol. **9**, (2008).

231. Katajamaa, M., Miettinen, J. & Orešič, M. MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* vol. **22**, 634–636 (2006).

232. Åberg, K. M., Torgrip, R. J. O., Kolmert, J., Schuppe-Koistinen, I. & Lindberg, J. Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of pure ion chromatograms using Kalman tracking. *J Chromatogr A* vol. **1192**, 139–146 (2008).

233. Zhu, H. *et al.* Feature Extraction for LC–MS via Hierarchical Density Clustering. *Chromatographia* vol. **82**, 1449–1457 (2019).

234. Reuschenbach, M., Hohrenk-Danzouma, L. L., Schmidt, T. C. & Renner, G. Development of a scoring parameter to characterize data quality of centroids in high-resolution mass spectra. *Anal Bioanal Chem* vol. **414**, 6635–6645 (2022).

235. Yu, H., Chen, Y. & Huan, T. Computational variation: An underinvestigated quantitative variability caused by automated data processing in untargeted metabolomics. *Anal Chem* vol. 93, 8719–8728 (2021) doi:10.1021/acs.analchem.0c03381.

236. Zhang, Z. *et al.* Reducing Quantitative Uncertainty Caused by Data Processing in Untargeted Metabolomics. *Anal Chem* vol. **96**, 3727–3732 (2024).

237. Guo, J. *et al.* EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained with over 25000 Extracted Ion Chromatograms. *Anal Chem* vol. **93**, 12181–12186 (2021).

238. Martens, H. Quantitative Big Data: where chemometrics can contribute. *J Chemom* vol. **29**, 563–581 (2015).

239. Gori, M. Learning Principles. in *Machine Learning* 60–121 (Elsevier, 2018). doi:10.1016/B978-0-08-100659-7.00002-6.

240. Harshman, R. A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-model factor analysis. *UCLA Working Papers in Phonetics* vol. 16, 1-84 (1970).

241. Harshman, R. A. PARAFAC2: Mathematical and technical notes. *UCLA Working Papers in Phonetics* vol. 22, 30-44 (1972)

242. Lawton, W. H. & Sylvestre, E. A. Self Modeling Curve Resolution. *Technometrics* vol. **13**, 617 (1971).

243. Tauler, R. Multivariate Curve Resolution Applied to Second Order Data. *Chemometrics and Intelligent Laboratory Systems*. vol. 30, 133-146 (1995).

244. Tauler, R. & Barcelo, D. Multivariate Curve Resolution Applied to Liquid Chromatography-Diode Array Detection. *TrAC - Trends in Analytical Chemistry* vol. 12 319-327 (1993)

245. Bro, R., Andersson, C. A. & Kiers, H. A. L. PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *J Chemom* vol. **13**, 295–309 (1999).

246. Kvalheim, O. M. & Liang, Y. *Characteristic Raman Frequencies of Organic Compounds*. *Holze, R. Electrochlm. Acta* vol. 64 https://pubs.acs.org/sharingguidelines (1992).

247. Windig, W., Liebman, S. A., Wasserman, M. B. & Peter Snyder, A. Fast Self-Modeling Curve-Resolution Method for Time-Resolved Mass Spectral Data. *Anal Chem* vol. 60, 1503-1510 https://pubs.acs.org/sharingguidelines (1988).

248. Sanchez, Eugenio. & Kowalski, B. R. Generalized rank annihilation factor analysis. *Anal Chem* vol. **58**, 496–499 (1986).

249. Windig, W., Heckler, C. E., Agblevor, F. A. & Evans, R. J. Self-Modeling Mixture Analysis of Categorized Pyrolysis Mass Spectral Data with the SIMPLISMA Approach. *Chemometrics and Intelligent Laboratory Systems* vol. 14, 195-207 (1992).

250. Jaumot, J., de Juan, A. & Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemometrics and Intelligent Laboratory Systems* vol. **140**, 1–12 (2015).

251. Daszykowski, M. & Walczak, B. Use and abuse of chemometrics in chromatography. *TrAC Trends in Analytical Chemistry* vol. **25**, 1081–1096 (2006).

252. Bos, T. S. *et al.* Recent applications of chemometrics in one- and two-dimensional chromatography. *J Sep Sci* vol. **43**, 1678–1727 (2020).

253. Niezen, L. E., Schoenmakers, P. J. & Pirok, B. W. J. Critical comparison of background correction algorithms used in chromatography. *Anal Chim Acta* vol. **1201**, 339605 (2022).

254. Risum, A. B. & Bro, R. Using deep learning to evaluate peaks in chromatographic data. *Talanta* vol. **204**, 255–260 (2019).

255. Baccolo, G., Quintanilla-Casas, B., Vichi, S., Augustijn, D. & Bro, R. From untargeted chemical profiling to peak tables – A fully automated AI driven approach to untargeted GC-MS. *TrAC - Trends in Analytical Chemistry* vol. 145 Preprint at https://doi.org/10.1016/j.trac.2021.116451 (2021).

256. Harshman, R. A. & Lundy, M. E. The PARAFAC model for three-way factor analysis and multidimensional scaling. in *Research methods for multimode data analysis* (eds. Law, H. G., Snyder Jr, C. W., Hattie, J. A. & McDonald, R. P.) 122–215 (Praeger, New York, 1984).

257. Jonsson, P. *et al.* High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal Chem* vol. **77**, 5635–5642 (2005).

258. Smirnov, A. *et al.* ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve Resolution to Spectral Deconvolution of Gas Chromatography-Mass Spectrometry Metabolomics Data. *Anal Chem* vol. **91**, 9069–9077 (2019).

259. Stein, S. E. *An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data*.

260. Pérez-López, C., Oró-Nolla, B., Lacorte, S. & Tauler, R. Regions of Interest Multivariate Curve Resolution Liquid Chromatography with Data-Independent Acquisition Tandem Mass Spectrometry. *Anal Chem* vol. **95**, 7519–7527 (2023).

261. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* vol. **84**, 283–289 (2012).

262. Tauler, R., Maeder, M. & de Juan, A. Multiset Data Analysis: Extended Multivariate Curve Resolution. in *Comprehensive Chemometrics* vol. 2 473–505 (Elsevier, 2009).

263. Olivieri, A. C. & Tauler, R. The effect of data matrix augmentation and constraints in extended multivariate curve resolution–alternating least squares. *J Chemom* vol. **31**, (2017).

264. Kiers, H. A. L., ten Berge, J. M. F. & Bro, R. PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *J Chemom* vol. **13**, 275–294 (1999).

265. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* vol. **38**, 149–171 (1997).

266. Zhang, X. & Tauler, R. Flexible Implementation of the Trilinearity Constraint in Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) of Chromatographic and Other Type of Data. *Molecules* vol. **27**, 2338 (2022).

267. Golub, G. H. & Reinsch, C. Singular value decomposition and least squares solutions. *Numer Math (Heidelb)* vol. **14**, 403–420 (1970).

268. Manne, R. On the Resolution Problem in Hyphenated Chromatography. *Chemometrics and Intelligent Laboratory Systems* vol. 27 (1995).

269. Johnsen, L. G., Amigo, J. M., Skov, T. & Bro, R. Automated resolution of overlapping peaks in chromatographic data. *J Chemom* vol. **28**, 71–82 (2014).

270. Kamstrup-Nielsen, M. H., Johnsen, L. G. & Bro, R. Core consistency diagnostic in PARAFAC2. *J Chemom* vol. **27**, 99–105 (2013).

271. Bro, R. & Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *J Chemom* vol. **17**, 274–286 (2003).

272. Fan, X. *et al.* Deep-Learning-Assisted multivariate curve resolution. *J Chromatogr A* vol. **1635**, 461713 (2021).

273. Sawall, M., Schröder, H., Meinhardt, D. & Neymeyr, K. 2.12 - On the Ambiguity Underlying Multivariate Curve Resolution Methods. in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Second Edition: Four Volume Set* vol. 2 199–231 (Elsevier, 2020).

274. Sawall, M. & Neymeyr, K. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: Theoretical foundation, inverse polygon inflation, and FAC-PACK implementation. *J Chemom* vol. **28**, 633–644 (2014).

275. Carabajal, M. D., Vidal, R. P., Arancibia, J. A. & Olivieri, A. C. A new constraint to model background signals when processing chromatographic-spectral second-order data with multivariate curve resolution. *Anal Chim Acta* vol. **1266**, (2023).

276. Tauler, R. and M. I. and C. E. Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths. *J Chemom* vol. **12**, 55–75 (1998).

277. Mishra, P. *et al.* Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC - Trends in Analytical Chemistry* vol. 137 Preprint at https://doi.org/10.1016/j.trac.2021.116206 (2021).

278. Smilde, A. K. *et al.* Common and distinct components in data fusion. *J Chemom* vol. **31**, (2017).

279. Lim, L.-H. Optimal solutions to non-negative PARAFAC/multilinear NMF always exist. *Workshop on Tensor Decomposition and Applications.* *https://www.stat.uchicago.edu/~lekheng/work/wtda1.pdf* , accessed 31.08.2024 (2005).

280. Bro, R. *Multiway Analysis in the Food Industry. Models, Algorithms and Applications*. https://www.researchgate.net/publication/2407346.

281. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM Review* vol. 51 455–500 Preprint at https://doi.org/10.1137/07070111X (2009).

282. Kruskal, J. B. Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics. *Linear Algebra and its Applications* vol: 18, 95-138 (1977)

283. Kruskal, J. B. Rank, decomposition, and uniqueness for 3-way and n -way arrays. *Multiway Data Analysis*, 7-18 (1989)

284. Hinrich, J. L., Madsen, K. H. & Mørup, M. The probabilistic tensor decomposition toolbox. *Mach Learn Sci Technol* vol. **1**, (2020).

285. Tomasi, G. & Bro, R. A comparison of algorithms for fitting the PARAFAC model. *Comput Stat Data Anal* vol. **50**, 1700–1734 (2006).

286. Bro, R., Andersson, C. A. & Kiers, H. A. L. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. *J Chemom* vol. **13**, 295–309 (1999).

287. Kiers, H. A. L. An Alternating Least Squares Algorithm for PARAFAC2 and Three-Way DEDICOM. *Computational Statistics & Data Analysis* vol. 16 (1993).

288. Cohen Jeremy E. and Bro, R. Nonnegative PARAFAC2: A Flexible Coupling Approach. in *Latent Variable Analysis and Signal Separation* (ed. Deville Yannick and Gannot, S. and M. R. and P. M. D. and W. D.) 89–98 (Springer International Publishing, Cham, 2018).

289. Stegeman, A. Degeneracy in candecomp/parafac and indscal explained for several three-sliced arrays with a two-valued typical rank. *Psychometrika* vol. **72**, 601–619 (2007).

290. Stegeman, A. Degeneracy in candecomp/parafac explained for p × p × 2 arrays of rank p + 1 or higher. *Psychometrika* vol. **71**, 483–501 (2006).

291. Lim, L. H. & Comon, P. Nonnegative approximations of nonnegative tensors. *J Chemom* vol. **23**, 432–441 (2009).

292. Rayens, W. S. & Mitchell, B. C. *Two-Factor Degeneracies and a Stabilization of PARAFAC Chemometrics and Intelligent Laboratory Systems*. *Chemometrics and Intelligent Laboratory Systems* vol. 38 (1997).

293. Mitchell, B. C. & Burdick, D. S. Slowly converging parafac sequences: Swamps and two-factor degeneracies. *J Chemom* vol. **8**, 155–168 (1994).

294. Yu, H., Bro, R. & Gallagher, N. B. PARASIAS: A new method for analyzing higher-order tensors with shifting profiles. *Anal Chim Acta* vol. **1238**, 339848 (2023).

295. Tauler, R. Multivariate curve resolution of multiway data using the multilinearity constraint. *J Chemom* vol. **35**, (2021).

296. Rood, D. B. B. Gas Chromatography Problem Solving and Troubleshooting. *J Chromatogr Sci* vol. **42**, 54–55 (2004).

297. Knepper, T. P. *et al.* Detection of polar organic substances relevant for drinking water. *Waste Management* vol. **19**, 77–99 (1999).

298. Bowden, J. A., Colosi, D. M., Mora-Montero, D. C., Garrett, T. J. & Yost, R. A. Evaluation of Derivatization Strategies for the Comprehensive Analysis of Endocrine Disrupting Compounds using GC/MS. *J Chromatogr Sci* vol. **47**, 44–51 (2009).

299. Hoggard, J. C. & Synovec, R. E. Parallel factor analysis (PARAFAC) of target analytes in GC × GC-TOFMS data: Automated selection of a model with an appropriate number of factors. *Anal Chem* vol. **79**, 1611–1619 (2007).

300. Hoggard, J. C. & Synovec, R. E. Automated resolution of nontarget analyte signals in GC x GC-TOFMS data using parallel factor analysis. *Anal Chem* vol. **80**, 6677–6688 (2008).

301. Hoggard, J. C., Siegler, W. C. & Synovec, R. E. Toward automated peak resolution in complete GC × GC-TOFMS chromatograms by PARAFAC. *J Chemom* vol. **23**, 421–431 (2009).

302. Pinkerton, D. K., Parsons, B. A., Anderson, T. J. & Synovec, R. E. Trilinearity deviation ratio: A new metric for chemometric analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data. *Anal Chim Acta* vol. **871**, 66–76 (2015).

303. Skov, T., Hoggard, J. C., Bro, R. & Synovec, R. E. Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling. *J Chromatogr A* vol. **1216**, 4020–4029 (2009).

304. Yaws, C. L. Enthalpy of Vaporization – Organic Compounds. in *Thermophysical Properties of Chemicals and Hydrocarbons* 366–489 (Elsevier, 2014). doi:10.1016/b978-0-323-28659-6.00007-0.

305. Parastar, H. *et al.* Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution. *Anal Chem* vol. **83**, 9289–9297 (2011).

306. Schneide, P. A., Bro, R. & Gallagher, N. B. Shift-invariant tri-linearity—A new model for resolving untargeted gas chromatography coupled mass spectrometry data. *J Chemom* **37**, (2023).

307. Rafferty, J. L., Zhang, L., Siepmann, J. I. & Schure, M. R. Retention mechanism in reversed-phase liquid chromatography: A molecular perspective. *Anal Chem* vol. **79**, 6551–6558 (2007).

308. Rigano, F. *et al.* The retention index approach in liquid chromatography: An historical review and recent advances. *J Chromatogr A* vol. **1640**, (2021).

309. Anzardi, M. B., Arancibia, J. A. & Olivieri, A. C. Interpretation of matrix chromatographic-spectral data modeling with parallel factor analysis 2 and multivariate curve resolution. *J Chromatogr A* vol. **1604**, (2019).

310. Khakimov, B., Amigo, J. M., Bak, S. & Engelsen, S. B. Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *J Chromatogr A* **1266**, 84–94 (2012).

311. Turova, P., Rodin, I., Shpigun, O. & Stavrianidi, A. A new PARAFAC-based algorithm for HPLC–MS data treatment: herbal extracts identification. *Phytochemical Analysis* vol. **31**, 948–956 (2020).

312. Farrés, M., Piña, B. & Tauler, R. Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS. *Metabolomics* vol. **11**, 210–224 (2015).

313. Navarro-Reig, M., Jaumot, J., García-Reiriz, A. & Tauler, R. Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies. *Anal Bioanal Chem* vol. **407**, 8835–8847 (2015).

314. Dalmau, N., Bedia, C. & Tauler, R. Validation of the Regions of Interest Multivariate Curve Resolution (ROIMCR) procedure for untargeted LC-MS lipidomic analysis. *Anal Chim Acta* vol. **1025**, 80–91 (2018).

315. Pérez-Cova, M., Platikanov, S., Tauler, R. & Jaumot, J. Quantification strategies for two-dimensional liquid chromatography datasets using regions of interest and multivariate curve resolution approaches. *Talanta* vol. **247**, (2022).

316. Navarro-Reig, M., Jaumot, J., van Beek, T. A., Vivó-Truyols, G. & Tauler, R. Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil. *Talanta* vol. **160**, 624–635 (2016).

317. Pérez-Cova, M., Tauler, R. & Jaumot, J. Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior. *Chemometrics and Intelligent Laboratory Systems* vol. **201**, (2020).

318. Mondello, L. *et al.* Comprehensive two-dimensional liquid chromatography. *Nature Reviews Methods Primers* vol. **3**, 86 (2023).

319. Fraga, C. G. & Corley, C. A. The chemometric resolution and quantification of overlapped peaks form comprehensive two-dimensional liquid chromatography. *J Chromatogr A* vol. **1096**, 40–49 (2005).

320. Bailey, H. P., Rutan, S. C. & Carr, P. W. Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis. *J Chromatogr A* vol. **1218**, 8411–8422 (2011).

321. Bayat, M., Marín-García, M., Ghasemi, J. B. & Tauler, R. Application of the area correlation constraint in the MCR-ALS quantitative analysis of complex mixture samples. *Anal Chim Acta* vol. **1113**, 52–65 (2020).

322. Vonk, R. J. *et al.* Comprehensive Two-Dimensional Liquid Chromatography with Stationary-Phase-Assisted Modulation Coupled to High-Resolution Mass Spectrometry Applied to Proteome Analysis of *Saccharomyces cerevisiae*. *Anal Chem* vol. **87**, 5387–5394 (2015).

323. Mahieu, N. G., Genenbacher, J. L. & Patti, G. J. A roadmap for the XCMS family of software solutions in metabolomics. *Current Opinion in Chemical Biology* vol. 30 87–93 Preprint at https://doi.org/10.1016/j.cbpa.2015.11.009 (2016).

324. Damiani, T. *et al.* Software and Computational Tools for LC-MS-Based Epilipidomics: Challenges and Solutions. *Analytical Chemistry* vol. 95 287–303 Preprint at https://doi.org/10.1021/acs.analchem.2c04406 (2023).

325. Hiroshi Tsugawa. *MS-DIAL Tutorial*. https://systemsomicslab.github.io/mtbinfo.github.io/MS-DIAL/tutorial.html#section-8-3-3 (2020).

326. Smith, C. A. *et al. Package 'xcms' - LC-MS and GC-MS Data Analysis*. https://bioconductor.org/packages/devel/bioc/manuals/xcms/man/xcms.pdf (2024).

327. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal Chem* vol. **89**, 8696–8703 (2017).

328. Du, P., Kibbe, W. A. & Lin, S. M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* vol. **22**, 2059–2065 (2006).

329. Guo, J. & Huan, T. Mechanistic Understanding of the Discrepancies between Common Peak Picking Algorithms in Liquid Chromatography–Mass Spectrometry-Based Metabolomics. *Anal Chem* vol. **95**, 5894–5902 (2023).

330. Tsugawa, H., Kanazawa, M., Ogiwara, A. & Arita, M. MRMPROBS suite for metabolomics using large-scale MRM assays. *Bioinformatics* vol. **30**, 2379–2380 (2014).

331. Savitzky, Abraham. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem* **36**, 1627–1639 (1964).

332. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* vol. **12**, 115–121 (2015).

333. Kuhl, C., Tautenhahn, R., Treutler, H. & Neumann, S. Package 'CAMERA' - Collection of Annotation Related Methods for Mass Spectrometry Data. https://www.bioconductor.org/packages/release/bioc/manuals/CAMERA/man/CAMERA.pdf , accessed 31.08.2024 (2024).

334. Lai, Z. *et al.* Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* vol. **15**, 53–56 (2018).

335. Tada, I. *et al.* Correlation-Based Deconvolution (CorrDec) to Generate High-Quality MS2 Spectra from Data-Independent Acquisition in Multisample Studies. *Anal Chem* vol. **92**, 11310–11317 (2020).

336. Smirnov, A., Jia, W., Walker, D. I., Jones, D. P. & Du, X. ADAP-GC 3.2: Graphical Software Tool for Efficient Spectral Deconvolution of Gas Chromatography-High-Resolution Mass Spectrometry Metabolomics Data. *J Proteome Res* vol. **17**, 470–478 (2018).

337. Wehrens, R. Analysis of GC-MS Metabolomics Data with MetaMS. https://www.bioconductor.org/packages/release/bioc/vignettes/metaMS/inst/doc/runGC.pdf , accessed 31.08.2024 (2024).

338. Jiang, W. *et al.* An automated data analysis pipeline for GC-TOF-MS metabonomics studies. *J Proteome Res* vol. **9**, 5974–5981 (2010).

339. Ni, Y. *et al.* ADAP-GC 2.0: Deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies. *Anal Chem* vol. **84**, 6619–6629 (2012).

340. Hiller, K. *et al.* MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis. *Anal Chem* vol. **81**, 3429–3439 (2009).

341. Dromey, R. G., Stefik, M. J., Rindfleisch, T. C. & Duffield, A. M. Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data. *Anal Chem* vol. **48**, 1368–1375 (1976).

342. Åberg, K. M., Alm, E. & Torgrip, R. J. O. The correspondence problem for metabonomics datasets. *Anal Bioanal Chem* vol. **394**, 151–162 (2009).

343. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief Bioinform* vol. **16**, 104–117 (2015).

344. Katajamaa, M. & Orešič, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* vol. **6**, 179 (2005).

345. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* vol. **11**, (2010).

346. Fischler, M. A. & Bolles, R. C. Random sample consensus. *Commun ACM* vol. **24**, 381–395 (1981).

347. Prince, J. T. & Marcotte, E. M. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal Chem* vol. **78**, 6140–6152 (2006).

348. Schmid, R. MZmine documentation - Join aligner. https://mzmine.github.io/mzmine_documentation/module_docs/align_join_aligner/join_aligner.html , accessed 31.08.2024 (2022).

349. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Acoust* vol. **26**, 43–49 (1978).

350. Wang, C. Po. & Isenhour, T. L. Time-warping algorithm applied to chromatographic peak matching gas chromatography/Fourier transform infrared/mass spectrometry. *Anal Chem* vol. **59**, 649–654 (1987).

351. Tomasi, G., van den Berg, F. & Andersson, C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemom* vol. **18**, 231–241 (2004).

352. Skov, T., Van Den Berg, F., Tomasi, G. & Bro, R. Automated alignment of chromatographic data. *J Chemom* vol. **20**, 484–497 (2006).

353. Kováts, E. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helv Chim Acta* vol. **41**, 1915–1932 (1958).

354. van Den Dool, H. & Dec. Kratz, P. A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography. *J Chromatogr A* vol. **11**, 463–471 (1963).

355. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* vol. **112**, 12549–12550 (2015).

356. Chaleckis, R., Meister, I., Zhang, P. & Wheelock, C. E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Current Opinion in Biotechnology* vol. 55 44–50 Preprint at https://doi.org/10.1016/j.copbio.2018.07.010 (2019).

357. Bittremieux, W., Wang, M. & Dorrestein, P. C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics* vol. 18 Preprint at https://doi.org/10.1007/s11306-022-01947-y (2022).

358. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem* vol. **88**, 7946–7958 (2016).

359. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* vol. **16**, 299–302 (2019).

360. Zhou, Z. *et al.* Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun* vol. **13**, (2022).

361. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* vol. 34 828–837 Preprint at https://doi.org/10.1038/nbt.3597 (2016).

362. Allen, F., Pon, A., Wilson, M., Greiner, R. & Wishart, D. CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res* vol. **42**, (2014).

363. Wallace, B. *et al.* Welcome to the Official Release of the NIST23 Mass Spectral Libraries. *ASMS 2023* https://chemdata.nist.gov/dokuwiki/lib/exe/fetch.php?media=chemdata:asms2023:nist_2023_release_presentation.pdf accessed 31.08.2024 (2023).

364. Kopka, J. *et al.* GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* vol. **21**, 1635–1638 (2005).

365. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* vol. **45**, 703–714 (2010).

366. Jarmusch, S. A., Van Der Hooft, J. J. J., Dorrestein, P. C. & Jarmusch, A. K. Advancements in capturing and mining mass spectrometry data are transforming natural products research. *Natural Product Reports* vol. 38 2066–2082 Preprint at https://doi.org/10.1039/d1np00040c (2021).

367. Bittremieux, W. *et al.* Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment for Discovery of Structurally Related Molecules. *Anal Chem* vol. 33 1733-1744 (2022)

368. Schmid, R. *et al.* Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun* vol. **12**, 3832 (2021).

369. Safizadeh, H. *et al.* Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions. *Journal of Chemical Information and Modeling* vol. 61 4156–4172 Preprint at https://doi.org/10.1021/acs.jcim.0c00993 (2021).

370. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A* vol **112**, 12580–12585 (2015).

371. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res* vol. **44**, D1202–D1213 (2016).

372. Bauer, C. A. & Grimme, S. How to Compute Electron Ionization Mass Spectra from First Principles. *Journal of Physical Chemistry A* vol. **120**, 3755–3766 (2016).

373. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).

374. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: A web-based platform to process untargeted metabolomic data. *Anal Chem* **84**, 5035–5039 (2012).

375. Tsugawa, H. *et al.* A lipidome atlas in MS-DIAL 4. *Nat Biotechnol* **38**, 1159–1163 (2020).

376. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem* **88**, 7946–7958 (2016).

377. Tsugawa, H. Advances in computational metabolomics and databases deepen the understanding of metabolisms. *Current Opinion in Biotechnology* vol. 54 10–17 Preprint at https://doi.org/10.1016/j.copbio.2018.01.008 (2018).

378. Marney, L. C. *et al.* Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data. *Talanta* vol. **115**, 887–895 (2013).

379. Furbo, S., Hansen, A. B., Skov, T. & Christensen, J. H. Pixel-based analysis of comprehensive two-dimensional gas chromatograms (color plots) of petroleum: A tutorial. *Anal Chem* vol. **86**, 7160–7170 (2014).

380. Johnson, K. J. & Synovec, R. E. Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems* vol. **60**, 225–237 (2002).

381. Hantao, L. W. *et al.* Comprehensive two-dimensional gas chromatography combined to multivariate data analysis for detection of disease-resistant clones of Eucalyptus. *Talanta* vol. **116**, 1079–1084 (2013).

382. Mohler, R. E. *et al.* Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software. *Analyst* vol. **132**, 756–767 (2007).

383. Poulsen, K. G., Kristensen, M., Tomasi, G., Cruz, M. Dela & Christensen, J. H. The Pixel-Based Chemometric Approach for Oil Spill Identification and Hydrocarbon Source Differentiation: Two Case Studies From the Persian Gulf. in *Oil Spill Environmental Forensics Case Studies* 443–463 (Elsevier, 2017). doi:10.1016/B978-0-12-804434-6.00021-5.

384. Christensen, J. H. & Tomasi, G. Practical aspects of chemometrics for oil spill fingerprinting. *Journal of Chromatography A* vol. 1169 1–22 Preprint at https://doi.org/10.1016/j.chroma.2007.08.077 (2007).

385. Nielsen, N. J., Ballabio, D., Tomasi, G., Todeschini, R. & Christensen, J. H. Chemometric analysis of gas chromatography with flame ionisation detection chromatograms: A novel method for classification of petroleum products. *J Chromatogr A* vol. **1238**, 121–127 (2012).

386. Christensen, J. H., Mortensen, J., Hansen, A. B. & Andersen, O. Chromatographic preprocessing of GC-MS data for analysis of complex chemical mixtures. *J Chromatogr A* vol. **1062**, 113–123 (2005).

387. Zhang, D., Huang, X., Regnier, F. E. & Zhang, M. Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data. *Anal Chem* vol. **80**, 2664–2671 (2008).

388. Tomasi, G., Savorani, F. & Engelsen, S. B. Icoshift: An effective tool for the alignment of chromatographic data. *J Chromatogr A* vol. **1218**, 7832–7840 (2011).

389. Vest Nielsen, N.-P., Carstensen, J. M. & Smedsgaard, J. *Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping*. *Journal of Chromatography A* vol. 805 www.imm.dtu.dk (1998).

390. Parsons, B. A. *et al.* Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC-TOFMS) Data Using a Null Distribution Approach. *Anal Chem* vol. **87**, 3812–3819 (2015).

391. Brownie, C., Boos, D. D. & Hughes-Oliver, J. Modifying the t and ANOVA F Tests When Treatment Is Expected to Increase Variability Relative to Controls. *Biometrics* vol. **46**, 259 (1990).

392. Prebihalo, S. E. *et al.* Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury. *Anal Chem* vol. **92**, 15526–15533 (2020).

393. Schöneich, S., Ochoa, G. S., Monzón, C. M. & Synovec, R. E. Minimum variance optimized Fisher ratio analysis of comprehensive two-dimensional gas chromatography / mass spectrometry data: Study of the pacu fish metabolome. *J Chromatogr A* vol. **1667**, (2022).

394. Sudol, P. E., Ochoa, G. S., Cain, C. N. & Synovec, R. E. Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. *Anal Chim Acta* vol. **1209**, (2022).

395. Lu, H., Liang, Y., Dunn, W. B., Shen, H. & Kell, D. B. Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. *TrAC - Trends in Analytical Chemistry* vol. **27**, 215–227 (2008).

396. Cerdán-Calero, M., Sendra, J. M. & Sentandreu, E. Gas chromatography coupled to mass spectrometry analysis of volatiles, sugars, organic acids and aminoacids in Valencia Late orange juice and reliability of the Automated Mass Spectral Deconvolution and Identification System for their automatic identification and quantification. *J Chromatogr A* vol. **1241**, 84–95 (2012).

397. Libiseller, G. *et al.* IPO: A tool for automated optimization of XCMS parameters. *BMC Bioinformatics* vol. **16**, (2015).

398. Albóniga, O. E., González, O., Alonso, R. M., Xu, Y. & Goodacre, R. Optimization of XCMS parameters for LC–MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. *Metabolomics* vol. **16**, (2020).

399. Lassen, J., Nielsen, K. L., Johannsen, M. & Villesen, P. Assessment of XCMS Optimization Methods with Machine-Learning Performance. *Anal Chem* vol. **93**, 13459–13466 (2021).

400. Guo, J., Shen, S. & Huan, T. Paramounter: Direct Measurement of Universal Parameters to Process Metabolomics Data in a 'white Box'. *Anal Chem* vol. **94**, 4260–4268 (2022).

401. Tauler, R., Gorrochategui, E., Jaumot, J. & Tauler, R. A protocol for LC-MS metabolomic data processing using chemometric tools. *Protoc Exch* (2015) doi:10.1038/protex.2015.102.

402. Yamamoto, F. Y. *et al.* Linking MS1 and MS2 signals in positive and negative modes of LC-HRMS in untargeted metabolomics using the ROIMCR approach. *Anal Bioanal Chem* vol. **415**, 6213–6225 (2023).

403. Queral-Beltran, A., Marín-García, M., Lacorte, S. & Tauler, R. UV-Vis absorption spectrophotometry and LC-DAD-MS-ESI(+)-ESI(−) coupled to chemometrics analysis of the monitoring of sulfamethoxazole degradation by chlorination, photodegradation, and chlorination/photodegradation. *Anal Chim Acta* vol. **1276**, (2023).

404. Sudol, P. E., Ochoa, G. S. & Synovec, R. E. Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *J Chromatogr A* vol. **1644**, 462092 (2021).

405. Guo, Z., Fan, Y., Yu, C., Lu, H. & Zhang, Z. GCMSFormer: A Fully Automatic Method for the Resolution of Overlapping Peaks in Gas Chromatography–Mass Spectrometry. *Anal Chem* vol. **96**, 5878–5886 (2024).

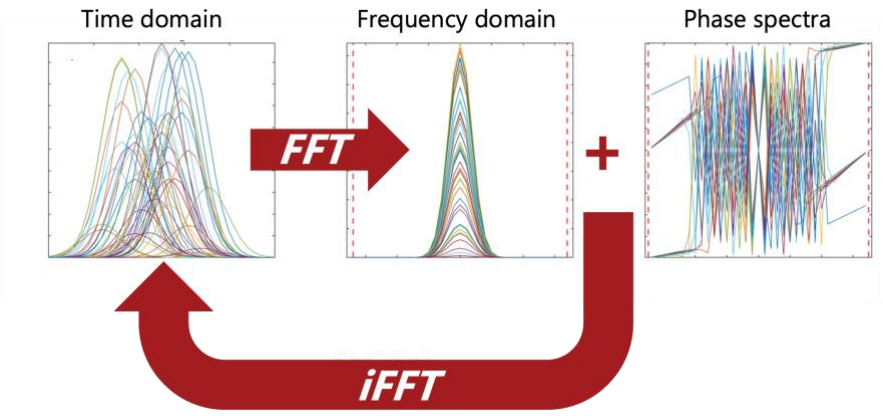
406. Fan, X. *et al.* Fully automatic resolution of untargeted GC-MS data with deep learning assistance. *Talanta* vol. **244**, (2022).

407. Fan, X. *et al.* Deep-Learning-Assisted multivariate curve resolution. *J Chromatogr A* vol. **1635**, (2021).

8 Research papers

Paper 1

Schneide P-A, Bro R, Gallagher NB. Shift-invariant tri-linearity—A new model for resolving untargeted gas chromatography coupled mass spectrometry data. Journal of Chemometrics. 2023; 37(8):e3501. doi:10.1002/cem.3501



Received: 24 January 2023 | Revised: 25 May 2023 | Accepted: 25 May 2023

DOI: 10.1002/cem.3501

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Shift-invariant tri-linearity—A new model for resolving untargeted gas chromatography coupled mass spectrometry data

Paul-Albert Schneide[1,2] | **Rasmus Bro**[1] | **Neal B. Gallagher**[3]

[1]Department of Food Science, University of Copenhagen, Frederiksberg, Denmark

[2]Analytical Science, BASF SE, Ludwigshafen am Rhein, Rhineland-Palatinate, Germany

[3]Eigenvector Research, Inc., Manson, Washington, USA

Correspondence
Rasmus Bro, Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg, Denmark.
Email: rb@life.ku.dk

Funding information
BASF SE

Abstract

Multi-way data analysis is popular in chemometrics for the decomposition of, for example, spectroscopic or chromatographic higher-order tensor datasets. Parallel factor analysis (PARAFAC) and its extension, PARAFAC2, are extensively employed methods in chemometrics. Applications of PARAFAC2 for untargeted data analysis of hyphenated gas chromatography coupled with mass spectrometric detection (GC-MS) have proven to be very successful. This is attributable to the ability of PARAFAC2 to account for retention time shifts and shape changes in chromatographic elution profiles. Despite its usefulness, the most common implementations of PARAFAC2 are considered quite slow. Furthermore, it is difficult to apply constraints (e.g., non-negativity) to the shifted mode in PARAFAC2 models. Both aspects are addressed by a new shift-invariant tri-linearity (SIT) algorithm proposed in this paper. It is shown on simulated and real GC-MS data that the SIT algorithm is 20–60 times faster than the latest PARAFAC2-alternating least squares (ALS) implementation and the PARAFAC2-flexible coupling algorithm. Further, the SIT method allows the implementation of constraints in all modes. Trials on real-world data indicate that the SIT algorithm compares well with alternatives. The new SIT method achieves better factor resolution than the benchmark in some cases and tends to need fewer latent variables to extract the same chemical information. Although SIT is not capable of modeling shape changes in elution profiles, trials on real-world data indicate the great robustness of the method even in those cases.

Abbreviations: ALS, alternating least squares; DFT, discrete Fourier transform; EFP, explained fit percentage; FFT, fast Fourier transform; GC-MS, gas chromatography coupled with mass spectrometric detection; MCR, multivariate curve resolution; PCA, principal component analysis; PC, principal component; PF2, PARAFAC2 (parallel factor analysis 2); PFFC, PARAFAC2-flexible coupling; S/N, signal-to-noise ratio; SIT, shift-invariant tri-linearity; SPME, solid-phase microextraction; TIC, total ion chromatogram.

1 | INTRODUCTION

Parallel factor analysis (PARAFAC), also known as canonical decomposition (CANDECOMP), is a tensor decomposition method that is used to find a unique low-rank approximation to higher-order tensor data.[1,2] An important assumption for the PARAFAC model to provide chemically meaningful models is that the data must follow a low-rank multilinear structure. Multi-linearity means that a sum of the outer products of a set of factors can approximate the data well. This concept is an extension of bi-linear modeling, as shown in Figure 1, and can also be seen as the mathematical model of the Beer–Lambert law. Examples of bi-linear models are multivariate curve resolution (MCR) or principal component analysis (PCA). Contrary to MCR, PARAFAC provides only one unique set of multi-linear factors that minimizes the residual error under reasonable assumptions.[3] Thus, the solution of the PARAFAC model is unique under mild conditions, as described by Harshman and Lundy[4] and proven by Kruskal.[5]

PARAFAC is often used in chemometrics to analyze datasets, such as excitation-emission fluorescence spectroscopy or hyphenated chromatography measurements.[6]

Unfortunately, the PARAFAC assumption of low-rank tri-linearity for chromatographic data is rarely fulfilled because of shifts of the peak signals in the retention time mode.[7,8] Therefore, the PARAFAC model is less useful for this type of data because a single latent variable cannot be used to model elution profiles across the sample mode (i.e., the principle of parallel proportional profiles).[7] A more flexible model that can handle the described deviations from tri-linearity is the PARAFAC2 model.[9,10] In the PARAFAC2 model, the factors in one mode are allowed to shift and change shape to some extent. More specifically, it is only required that the factors in this mode have constant cross-products. Thus, factors (elution profiles) in one mode can vary for each sample in another mode under the constraint that the cross-products between the varying factors remain constant from sample to sample. The PARAFAC2 model has been proven to be very successful in modeling hyphenated chromatographic data, like gas chromatography coupled with mass spectrometric detection (GC-MS).[8,11] However, some challenges still need to be addressed to facilitate the use of PARAFAC2 and broaden the application area.

In the classical alternating least squares (ALS) implementation of the PARAFAC2 model, it is not generally possible to impose constraints on the shifted mode.[10,12] This is because the constraints must be imposed on the product of two matrices and not on a single-factor matrix. Alternative implementations have been developed recently that enable the use of constraints in all modes.[12–15] Some of the developed approaches show unexpected properties (e.g., more components are required to achieve a suitable PARAFAC2 solution).[15] Although some important improvements have been made to increase computational efficiency,[16] most algorithms for calculating the PARAFAC2 model[13,16,17] are still comparatively slow. This is particularly problematic in the analysis of untargeted chromatographic data because hundreds, or even thousands, of low-rank PARAFAC2 models must be fitted to extract chemical information from a high-rank dataset.[11] This paper proposes an alternative algorithm for calculating the PARAFAC2 model that is fast and allows all modes to be constrained. It is based on a novel shift-invariant tri-linearity (SIT) constraint to achieve a unique solution. The SIT method conceptually relies on PARASIAS[15] and a flexible tri-linearity implementation.[18] It specifically overcomes the issue of rank inflation reported for the PARASIAS method.[15]

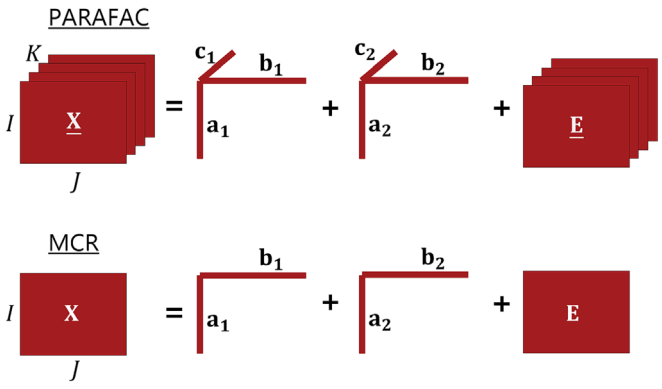


FIGURE 1 Parallel factor analysis (PARAFAC) as a tri-linear extension of bi-linear models like multivariate curve resolution (MCR).

The paper is organized as follows: Section 2 gives an overview of the relevant theoretical concepts leading to the introduction of the proposed method. Section 3 describes the experimental setup and the datasets used for the performance evaluation, and Section 4 provides the results accordingly. A conclusion is given in Section 5.

2 | THEORETICAL BACKGROUND

2.1 | PARAFAC and PARAFAC2

Let the matrix $\mathbf{X}_k$ $(I \times J)$ be the k^{th} slab of the tensor $\underline{\mathbf{X}}$ $(I \times J \times K)$. Then the rank R PARAFAC model is given by Equation (1) and the least squares loss function by Equation (2). In this model, $\mathbf{A}$ and $\mathbf{B}$ are factor matrices with dimensions $(I \times R)$ and $(J \times R)$ and $\mathbf{D}_k$ is a diagonal matrix with dimensions $(R \times R)$, containing the weights for the k^{th} slab. The diagonal elements of $\mathbf{D}_k$ are the elements of the k^{th} row of the factor matrix $\mathbf{C}$ $(K \times R)$ and $\mathbf{E}_k$ are the residuals with dimension $(I \times J)$.

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}} + \mathbf{E}_k \tag{1}$$

$$\underset{\mathbf{A},\mathbf{D},\mathbf{B}}{\operatorname{argmin}} \sum_k \left\|\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\right\|_F^2 \tag{2}$$

Algorithmically, the solutions for $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ can be found by ALS. In ALS, each factor matrix is estimated by solving a least squares problem, holding the other two factor matrices constant. The expression $\mathbf{B}\odot\mathbf{C}$ denotes the Kathri–Rao product (column-wise Kronecker product) of the matrices $\mathbf{B}$ and $\mathbf{C}$.

Update $\mathbf{A}$, holding $\mathbf{B}$ and $\mathbf{C}$ constant

$$\mathbf{A} = \mathbf{X}_{BC}(\mathbf{B}\odot\mathbf{C})\left((\mathbf{B}\odot\mathbf{C})^{\mathrm{T}}(\mathbf{B}\odot\mathbf{C})\right)^{-1}$$

Update $\mathbf{B}$, holding $\mathbf{A}$ and $\mathbf{C}$ constant

$$\mathbf{B} = \mathbf{X}_{AC}(\mathbf{C}\odot\mathbf{A})\left((\mathbf{C}\odot\mathbf{A})^{\mathrm{T}}(\mathbf{C}\odot\mathbf{A})\right)^{-1}$$

Update $\mathbf{C}$, holding $\mathbf{A}$ and $\mathbf{B}$ constant

$$\mathbf{C} = \mathbf{X}_{AB}(\mathbf{A}\odot\mathbf{B})\left((\mathbf{A}\odot\mathbf{B})^{\mathrm{T}}(\mathbf{A}\odot\mathbf{B})\right)^{-1}$$

For each update, $\underline{\mathbf{X}}$ must be unfolded along a different mode. The resulting matrices are visualized in Figure 2. Several computational adjustments to the procedure described above have been published to make the algorithm more efficient. A good summary is, for example, given by Tomasi and Bro.[19]

This procedure is repeated until the relative change in the loss function value is below a pre-defined convergence criterion.

The PARAFAC2 model was first described by Harshman.[9] It is an extension of the PARAFAC model that allows us to find for each k^{th} slab of $\underline{\mathbf{X}}$ an individual set of factors $\mathbf{B}_k$. Thus, the rank R PARAFAC2 model for the k^{th} slab of $\underline{\mathbf{X}}$ is given by Equation (3) and the loss function by Equation (4).

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k{}^{\mathrm{T}} + \mathbf{E}_k \tag{3}$$

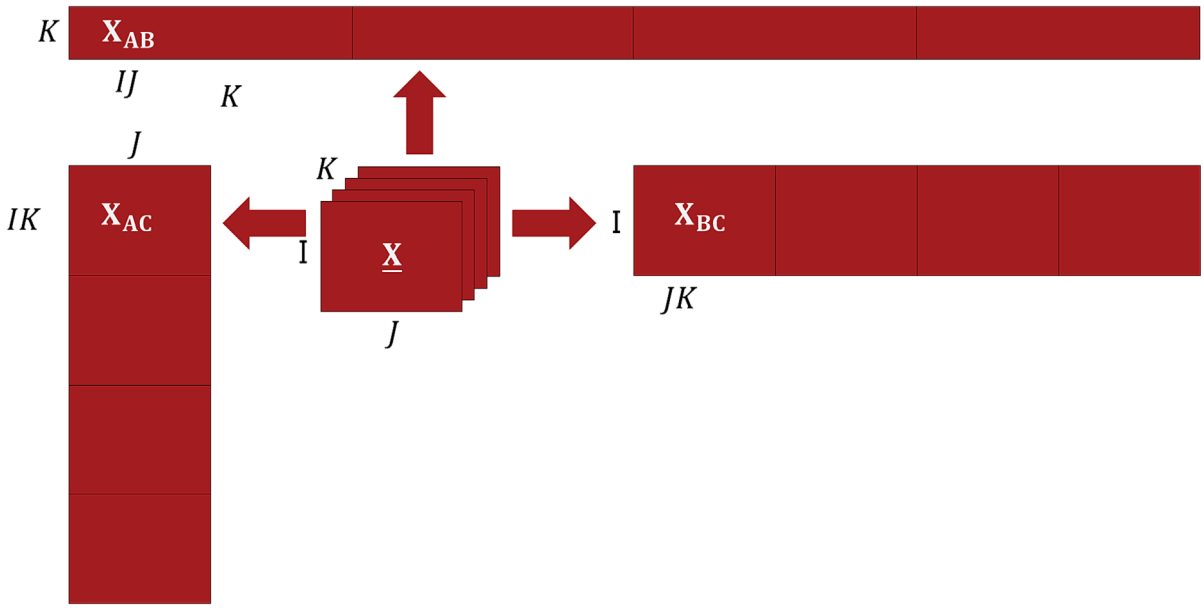


FIGURE 2 Different unfolding of a three-way array $\underline{\mathbf{X}}$.

$$\underset{\mathbf{A},\mathbf{D},\mathbf{B}}{\operatorname{argmin}} \sum_{k} \left\| \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}_k^{\mathrm{T}} \right\|_F^2 \tag{4}$$

To maintain uniqueness, the factor matrices $\mathbf{B}_k$ are constrained to have a constant cross-product matrix $\mathbf{H}$ for all k. This implies that $\mathbf{B}_k$ can be expressed as the matrix product of an orthonormal matrix $\mathbf{P}_k$ with the cross-product matrix $\mathbf{H}$ (cf. Equation (5)).

$$\mathbf{H} = \mathbf{B}_k^{\mathrm{T}}\mathbf{B}_k, \text{ for all } k = 1, \ldots, K \text{ and } \mathbf{B}_k = \mathbf{P}_k\boldsymbol{H} \tag{5}$$

The first efficient algorithm for fitting the PARAFAC2 model is the direct fitting algorithm proposed by Kiers et al.[10] Basically, the idea is to split the optimization problem into two steps. First, $\mathbf{P}_k$ is updated by calculating the SVD solution of the matrix $\left(\mathbf{X}_k^{\mathrm{T}}\mathbf{A}\mathbf{D}_k\mathbf{H}\right)$. In the second step, $\mathbf{X}_k$ is compressed by the projection on $\mathbf{P}_k$ and $\mathbf{A}$, and $\mathbf{D}_k$ and $\mathbf{H}$ are updated by running a few iterations of a PARAFAC model. However, one problem with this algorithm is that the $\mathbf{B}_k$ matrices cannot be constrained easily. This is because $\mathbf{B}_k$ is calculated implicitly by the product of the cross-product matrix $\mathbf{H}$ with the orthonormal matrix $\mathbf{P}_k$, as shown in Equation (5). The difficulties of constraining $\mathbf{B}_k$ in the classical PARAFAC2-ALS framework are described in detail by Cohen et al.[12]

More recently, the PARAFAC2-flexible coupling (PFFC) algorithm has been published that allows $\mathbf{B}_k$ to be, for example, non-negatively constrained. In this algorithm, the hard coupling $\mathbf{B}_k = \mathbf{P}_k\mathbf{H}$ is replaced by the more flexible coupling $\operatorname{argmin} \mu_k ||\mathbf{B}_k - \mathbf{P}_k\mathbf{H}||^2$. In other words, $\mathbf{B}_k$ must not be equal to but close to $\mathbf{P}_k\mathbf{H}$. The non-negative flexible coupling algorithm has been reported to be more robust and sometimes provides better factor resolution than the classical PARAFAC2-ALS.[12]

2.2 | Tri-linearity constraints

An alternative algorithm for calculating the PARAFAC model was proposed by Tauler.[20,21] In this algorithm, the loss function is formulated like an MCR problem, as shown in Equations (6) and (7). This approach is also referred to as multi-set analysis.[22] The matrix $\mathbf{X}_{BC}$ has dimensions $(IxJK)$ and can be obtained by unfolding $\underline{\mathbf{X}}$, as shown in Figure 2.

$$\mathbf{X}_{BC}^{\mathrm{T}} = \mathbf{V}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{6}$$

$$\underset{\mathbf{V},\mathbf{S}}{\operatorname{argmin}} \left\| \mathbf{X}_{BC}^{\mathrm{T}} - \mathbf{V}\mathbf{S}^{\mathrm{T}} \right\|_{F}^{2} \tag{7}$$

Deviating from the conventional notation in MCR, we use $\mathbf{V}$ instead of $\mathbf{C}$ to describe the matrix of concentration profiles to prevent confusion with the factor matrix $\mathbf{C}$ introduced in the PARAFAC notation. The factor matrix $\mathbf{V}$ has size $(JK \times R)$ and the factor matrix $\mathbf{S}$ is of size $(I \times R)$. To obtain a unique solution, a tri-linearity constraint is sequentially imposed on the reshaped columns of the factor matrix $\mathbf{V}$. Each column $\mathbf{v}_r$ is first reshaped into a $(J \times K)$ matrix $\mathbf{M}$. The tri-linearity assumption requires that the rows of $\mathbf{M}$ contain scalar multiples of the same factor. If tri-linearity holds, $\mathbf{M}$ should be approximately rank one at convergence up to noise. This is enforced by calculating $\mathbf{M}_{est}$ from the first principal component (PC) of $\mathbf{M}$ as an estimate of $\mathbf{M}$. A visualization of how the tri-linearity constraint is applied to $\mathbf{V}$ is given in Figure 3.

One advantage of the tri-linearity approach is that constraints can easily be applied component-wise.[20–23] Because tri-linearity is imposed factor-by-factor, another advantage is that tri-linearity can be relaxed factor-by-factor. This allows for greater flexibility in modeling deviations from tri-linearity with the potential expense of losing model uniqueness. Extensions of the tri-linearity constraint to quadri-linearity and multi-linearity have also been described.[24]

2.3 | Discrete Fourier transformation (DFT)

The DFT is one of the most important mathematical operations in signal processing and engineering. The DFT projects a discrete periodic signal with a period length N from the time domain into the frequency domain, as shown by Equation (8). The result is the discrete, complex-valued function $\widehat{f}[k]$, whose real part describes the amplitude of frequencies contained in the signal $f[n]$. The imaginary part of $\widehat{f}[k]$ describes the phase of the frequencies in $f[n]$. The inverse DFT reconstructs the signal in the time domain by a sum of weighted and phase-adjusted frequencies shown in Equation (9).

$$\widehat{\mathrm{f}}[\mathrm{k}] = \frac{1}{\mathrm{N}} \sum_{\mathrm{n}=0}^{\mathrm{N}-1} \mathrm{f}[\mathrm{n}]\, \mathrm{e}^{-\frac{2\pi \mathrm{i}}{\mathrm{N}}\mathrm{nk}}, \mathrm{k} = 0, 1, 2, \ldots, \mathrm{N}-1 \tag{8}$$

$$\mathrm{f}[\mathrm{n}] = \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=0}^{\mathrm{N}-1} \widehat{\mathrm{f}}[\mathrm{k}]\, \mathrm{e}^{\frac{2\pi \mathrm{i}}{\mathrm{N}}\mathrm{nk}}, \mathrm{n} = 0, 1, 2, \ldots, \mathrm{N}-1 \tag{9}$$

The power spectrum is defined as the squared modulus of $\widehat{f}$ as shown by Equation (10). One important property of the power spectrum is its shift invariance. The shift information of an input signal is contained in the phase spectrum.

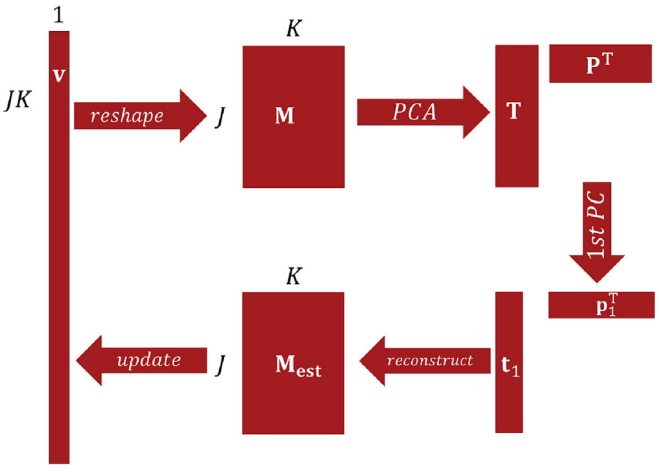


FIGURE 3 Visualization of how tri-linearity constraint can be imposed on the loading matrix $\mathbf{V}$.

$$\left\|\widehat{\mathbf{f}}\right\|_{\mathrm{F}}^{2}=\widehat{\mathbf{f}}\widehat{\mathbf{f}}^{*} \tag{10}$$

The shift-invariance property of the power spectrum is displayed in Figure 4. In this figure, the results of applying the DFT to a set of signals are illustrated. In (A), this set of signals was created by multiplying random weights to a Gaussian peak, and in (D), the same signal has been shifted randomly along the time axis. As can be seen by comparing (B) and (E), the shift in the time domain does not change the power spectrum of the signal. The shift information is, however, captured in the phase spectrum (F).

In practice, the DFT is calculated by the fast Fourier transformation (FFT) algorithm.[25] Using the most common radix-2 FFT algorithms requires the length of the signal to be a power of two. This can be easily achieved using zero-padding to convert the input signal length to the next larger power of two. A detailed description of other important properties of the DFT can, for example, be found in engineering or signal processing textbooks.[26]

2.4 | SIT

One flexible implementation of the tri-linearity approach for handling shifted data was recently described for chromatographic data.[18] In this implementation, profiles are aligned based on some specific points within the signal (e.g., the peak maximum). However, the alignment procedure must be formulated explicitly, which limits the applicability of this approach.

To illustrate the difference between explicit and implicit alignment, we consider three signals with arbitrary intensity (Figure 5). The first one is a normal Gaussian shape, and the second and third are exponentially modified Gaussian shapes (known to be a good model for skewed peaks in chromatography).[27] If the three signals are aligned based on the peak maximum, the result is a rank three system. Conversely, if the same three signals are aligned to a common phase spectrum, the result is a rank two system. The reason for this difference is that the shift-invariant power spectra of the

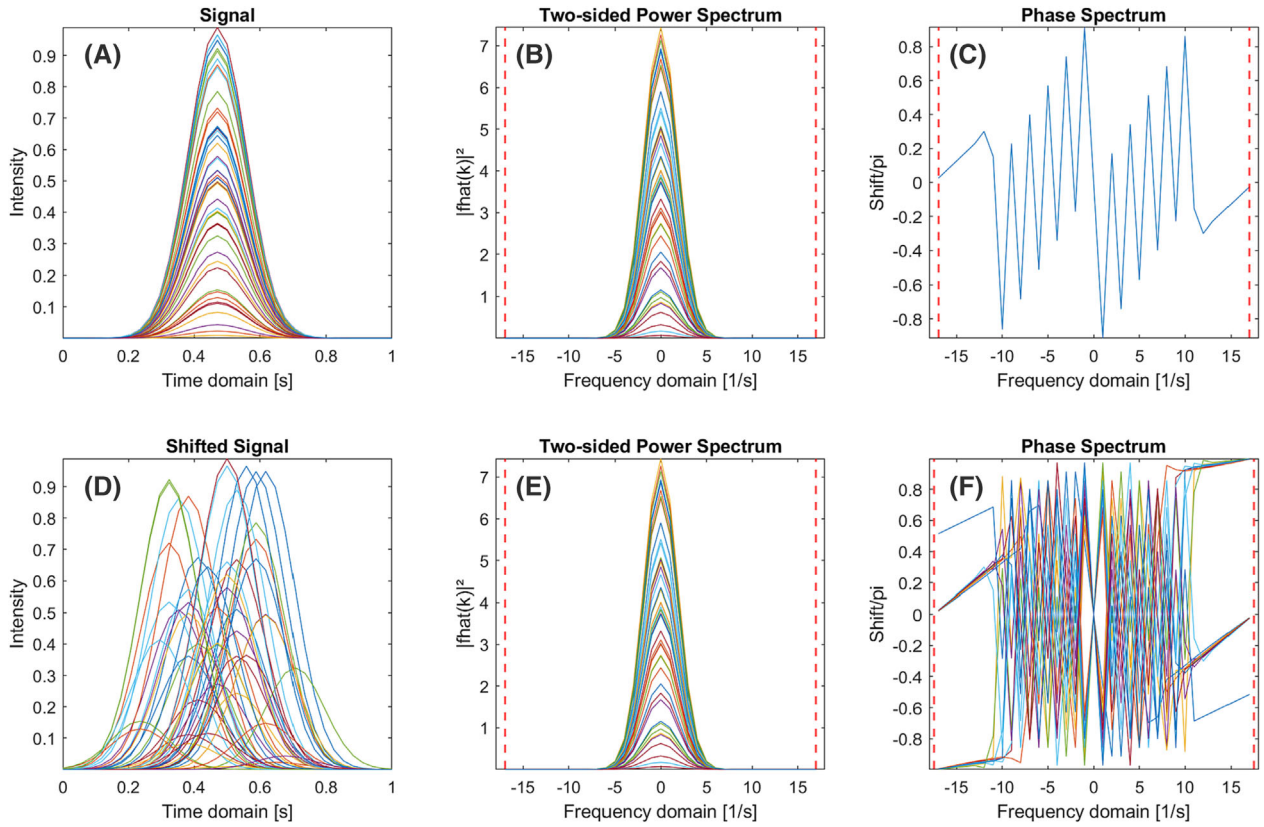


FIGURE 4 The Fourier transform of a set of differently scaled Gaussian profiles (A) and (D) is shown. Profiles in (D) have been shifted randomly along the time domain. The power spectra of (A) and (D) show no differences, as can be seen from (B) and (E). The shift information is captured in the phase spectra (C) and (F).

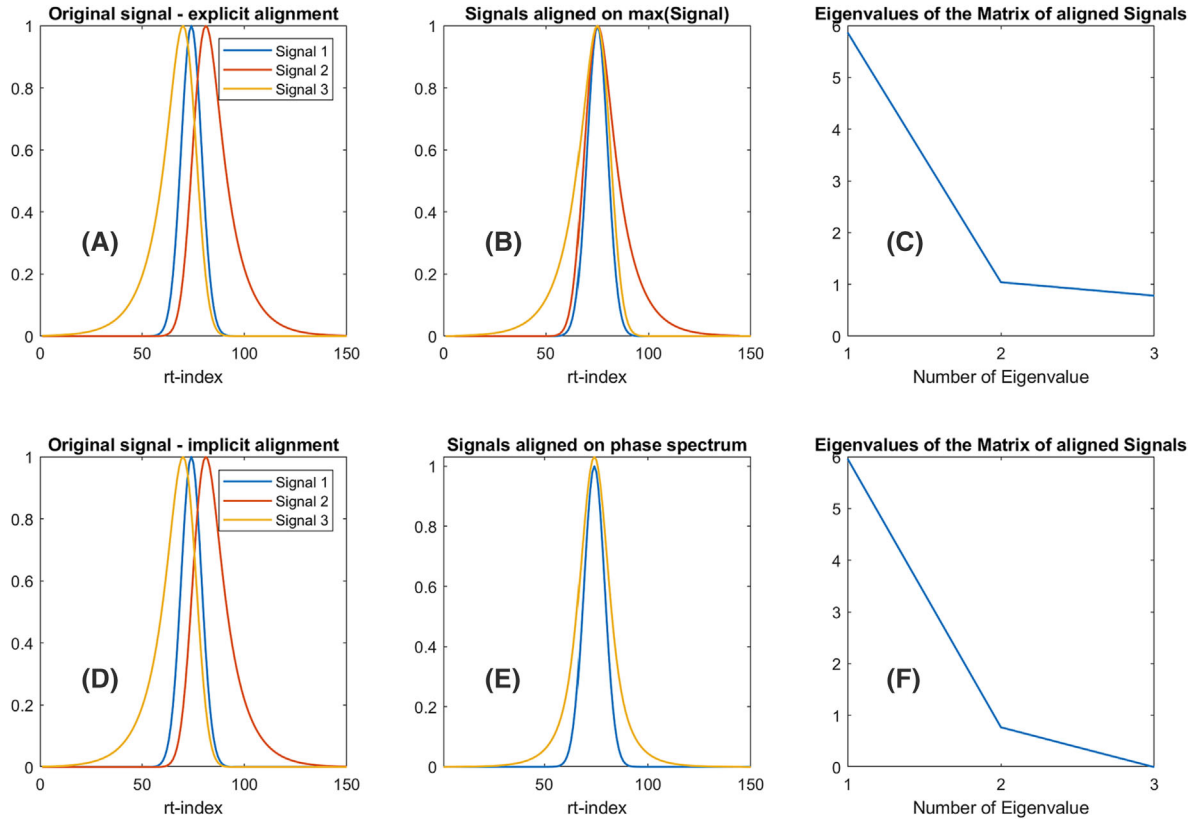


FIGURE 5 Illustration of the differences between implicit and explicit alignment. The top row (A–C) shows what happens to the original signals after an explicit alignment using the peak maximum as a fix point. The lower row (D–F) shows what happens to the same signals applying an implicit alignment based on a common phase spectrum. In summary, the rank of the aligned signals (B) shown in (C) is three, whereas the rank of the aligned signals (E) shown in (F) is two.

second and third signals are identical and differ only in their phase. Hence, it can be concluded that an implicit alignment is more broadly applicable than an explicit alignment.

A method for modeling shifted GC-MS data implicitly is PARASIAS.[15] In this method, the elution profiles are transformed into shift-invariant power spectra using the DFT. The power spectra are then decomposed using the PARAFAC algorithm. Finally, the inverse DFT is applied to the factor matrix containing the estimates of the elution profiles in the frequency domain. A significant drawback of the PARASIAS decomposition is that it sometimes requires more latent variables than PARAFAC2 to achieve comparable results on the same dataset.[15] This is because the rank of the power spectra matrix is inflated when the raw elution profiles are a sum of multiple underlying elution profiles. This is a consequence of the presence of cross-terms in the power spectra of the raw profiles. In contrast, the proposed SIT is applied factor-by-factor and does not introduce cross-terms. This means that SIT does not artificially increase the rank, as does PARASIAS.

The proposed SIT method combines the idea of the flexible tri-linearity approach[18] and the idea of using DFT for the synchronization of shifted profiles[15] and is illustrated in Figure 6. The DFT is used to synchronize the shifted profiles stored in the reshaped factor matrix $\mathbf{M}$ (cf. Figures 3 and 4). The DFT is computed column-wise on each of the K profiles in $\mathbf{M}$. The profiles are zero-padded as needed before calculating the DFT. Zero-padding can be used to adjust for differences in the length of individual profiles before reshaping $\mathbf{v}$ to $\mathbf{M}$. After applying the DFT to the columns of $\mathbf{M}$, a $J \times K$ matrix $\boldsymbol{\phi}$ containing the phase spectra and $J \times K$ matrix containing the shift-invariant power spectra $\widehat{\mathbf{M}}$ are obtained. The power spectra are calculated using Equation (10). Tri-linearity is imposed by performing a PCA decomposition of $\widehat{\mathbf{M}}$ and calculating $\widehat{\mathbf{M}}_{est}$ as the outer product of the first PC. The estimate of the original, shifted signal in time domain $\mathbf{M}_{est}$ is obtained by applying the inverse DFT to $\boldsymbol{\phi}$ and $\widehat{\mathbf{M}}_{est}$.

The proposed method will provide an accurate solution if the matrix of power spectra $\widehat{\mathbf{M}}$ has rank one for all R components. It is important to note that non-negativity on $\mathbf{V}$ is applied when $\mathbf{V}$ is estimated as the non-negative least squares solution of $\mathbf{X}_{BC}^{\mathrm{T}}\mathbf{S}\left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}$ (compared with Equation (6)). Subsequently, factors can still become slightly negative after calculating the inverse DFT of $\widehat{\mathbf{M}}_{est}$ to obtain $\mathbf{M}_{est}$. This can be prevented by setting negative entries in $\mathbf{M}_{est}$ to

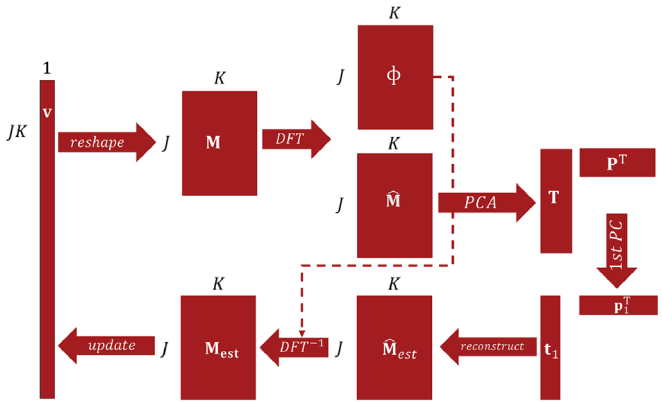


FIGURE 6 Shift-invariant tri-linearity (SIT) is based on a domain transformation followed by performing a principal component analysis (PCA) on the power spectra of the elution profiles matrix $\widehat{\boldsymbol{M}}$, rather than on $\boldsymbol{M}$ directly. Zero-padding might be used to adjust the profile length J to a power of two or to correct for differences in the profile length between individual profiles.

zero. The implementation of SIT shown in Figure 6 is not capable of modeling shifted, non-tri-linear data (e.g., elution profiles having different shapes across samples). In this situation, the rank of $\widehat{\mathbf{M}}$ will be greater than one. This limitation can potentially be overcome by softening the SIT constraint. One approach to do this is by taking more than the first PC for the calculation of $\widehat{\mathbf{M}}_{est}$. Thus, factors not behaving in a shifted, tri-linear manner can be modeled as bi-linear factors.[22,23] This approach allows for great flexibility, as has been shown in previous studies.[22,23] However, relaxing tri-linearity constraints on some factors is expected to affect the uniqueness properties of the solution.[28] A detailed description of the algorithm is provided in Appendix A.

3 | EXPERIMENTAL

3.1 | Algorithms and calculations

Three different algorithms were compared: (1) PARAFAC2-ALS (PF2), (2) PFFC (both implemented in the PARAFAC2-based Deconvolution and Identification System [PARADISe][29] toolbox), and (3) SIT. To make it a fair comparison, simple random initialization was selected throughout all experiments. Non-negativity constraints were applied to the non-shifted modes for (1) and to all modes for (2) and (3). The convergence criterion was selected as 10^{-9} for the relative change in the loss function. This means that the sequence would stop as soon as the relative change in loss function error between consecutive iterations is below 10^{-9}. Further, the criterion to distinguish between local and global minima was selected as $\Delta\text{EFP} < 10^{-8}$, in accordance with the heuristic described by Yu et al.[30] In this case, ΔEFP means the absolute difference in explained fit percentage between the best fit model $m_{1,R}$ and any other model $m_{i,R}$ taken from an ordered set of models with R components calculated on a specific dataset $\mathbf{X}$. The maximum number of iterations was set to 2×10^4.

All calculations and data analysis were conducted in MATLAB® R2020b 64-bit (The MathWorks, Inc., Natick, MA, USA) running 64-bit Microsoft Windows 10 Enterprise on an HP ZBook Fury G7 with an Intel® Core™ i9-10885H 2.40GHz CPU and 32 GB RAM.

3.2 | Datasets

3.2.1 | Simulated data

GC-MS datasets were simulated using the tool described in Tian et al.,[17] which is available and documented under https://ucphchemometrics.com/simulated-gc-ms-data/ [December 15, 2022]. To test the performance of the algorithms under different conditions, the number of components and the noise level of the simulated datasets were varied on two

levels, which gives four different conditions for a fully crossed design. The dimensions of the simulated factor matrices were $\mathbf{A}\,(200 \times R)$, $\mathbf{B}_k\,(70 \times R)$, and $\mathbf{C}\,(50 \times R)$, with R being the number of components, varied according to Table 1.

For each condition, 50 different $70 \times 50 \times 200$ sized tensors were simulated. Data examples for each condition are shown as total ion chromatograms (TICs) in Figure 7 together with the true underlying elution profiles. Each of these 200 (50 different for each of the four conditions) tensors was fitted ten times with each algorithm, starting from different random initializations. The models with the highest explained fit percentage (EFP) out of the ten repetitive fits were then compared. This procedure was used to avoid comparisons biased by the presence of local minima and degenerate solutions. The goal of this experiment was to compare the performance of the different algorithms in terms of computation time, lack of fit, and quality of factor resolution.

3.2.2 | Apple wine

Solid-phase microextraction (SPME) GC-MS measurements have been carried out to monitor the fermentation of apple wine with three different yeast species. A total of 155 samples have been measured over 12 days of fermentation. During the runtime of 46 min, 7091 mass scans from 15 to 300 m/z were recorded to determine the volatiles formed during the fermentation process. The dataset was collected as part of a master's thesis at the Department of Food Science at the University of

TABLE 1 Showing the selected conditions for the simulation of different gas chromatography coupled with mass spectrometric detection (GC-MS) datasets.

Condition	Number of components	Signal-to-noise ratio
1	3	100
2	3	4
3	5	100
4	5	4

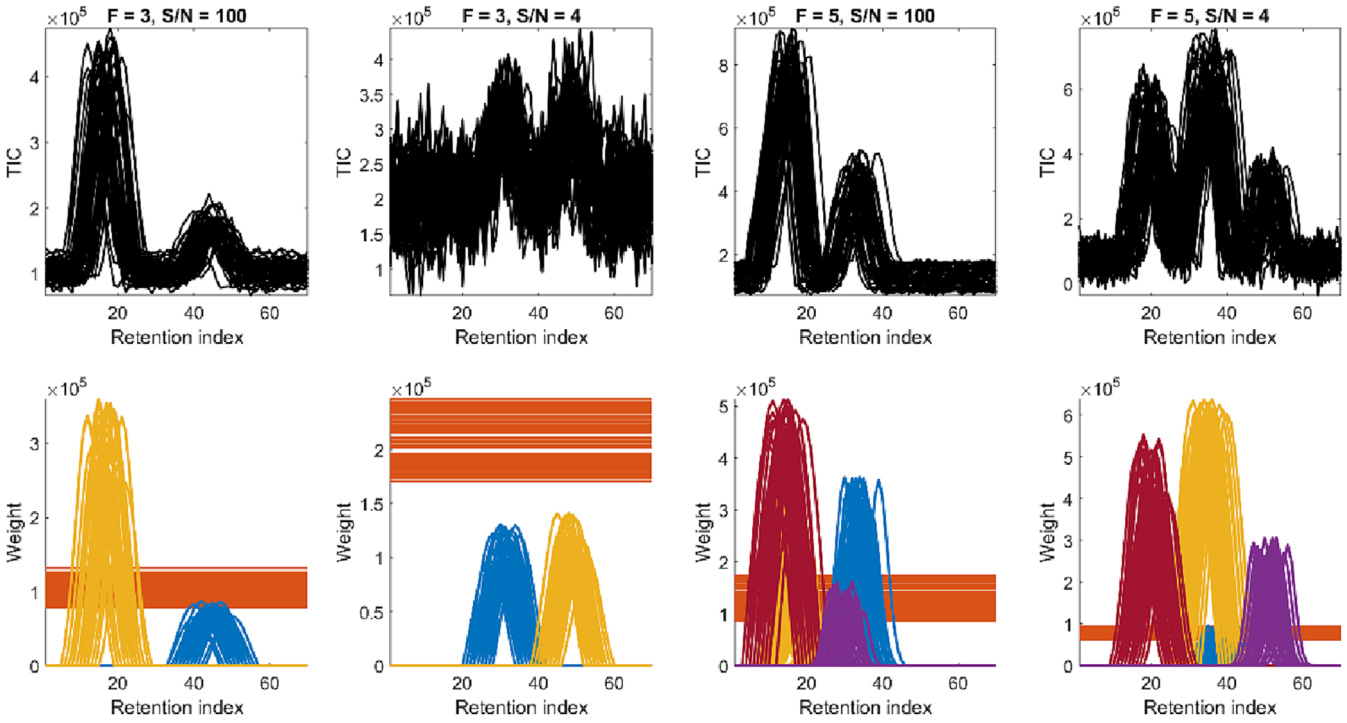


FIGURE 7 Top row shows examples of the simulated datasets (total ion chromatogram [TIC] profiles) for the conditions defined in Table 1. Bottom row shows the true underlying elution profiles multiplied by the relative concentrations. Similarly, 49 additional datasets have been simulated and modeled with shift-invariant tri-linearity (SIT), PARAFAC2–ALS (PF2), and PARAFAC2–flexible coupling (PFFC).

Copenhagen. The results are not published. Three intervals were selected that cover different real challenges in the analysis of GC-MS data. More specifically, Interval 12 has a peak with a low signal-to-noise ratio (S/N). Interval 18 contains one large, shifting peak that changes its shape with increasing intensity. Interval 33 contains two small peaks in the tail of one large cut-off peak. The selected datasets are shown in Figure 8. The goal of this experiment is to qualitatively evaluate how the different algorithms cope with difficult real-world data and to study the limitations of all algorithms comparatively.

4 | RESULTS

4.1 | Simulated data

The results of the benchmark study on simulated data are shown in Table 2. Considering all the investigated conditions, SIT is drastically faster than the benchmark. Both PF2 and PFFC show overall similar performance but are

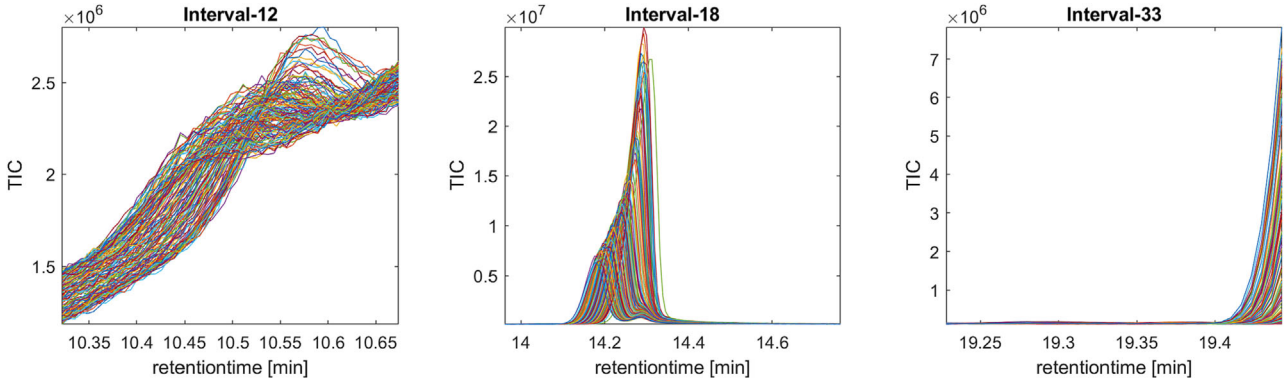


FIGURE 8 Gas chromatography coupled with mass spectrometric detection (GC-MS) data measured on different batches of apple wine fermentation. Interval 12: large baseline component and small peak. Interval 18: intense peak that changes its shape with increasing intensity. Interval 33: two very small peaks in the tail of a very large peak.

TABLE 2 Results for the benchmark study on simulated data. The results in columns "Iterations", "Computation time", and "EFP_{max}" were calculated as the mean over the 50 datasets for each respective condition. The "Tucker congruence" has been calculated by averaging first over the Tucker congruences of the factor matrices for each of the three modes in one dataset and then by taking the mean over all 50 datasets.

							Tucker congruence		
Algorithm	**Conds.**	**Comps.**	**EFP_{max} (%)**	**EFP_{dif} (%)**	**Iterations**	**Computation time (s)**	**A**	**B**	**C**
PF2	1	3	90.054	0.009	5542	24.0	1.000	1.000	1.000
PFFC	1	3	90.054	0.013	1549	20.0	0.979	0.989	1.000
SIT	1	3	90.054	0.008	139	0.7	1.000	1.000	1.000
PF2	2	3	80.880	0.295	3779	14.0	0.976	0.989	0.999
PFFC	2	3	80.880	0.249	809	11.0	0.973	0.984	1.000
SIT	2	3	80.880	0.173	109	0.5	1.000	0.997	1.000
PF2	3	5	99.061	0.022	3657	42.0	0.999	0.999	1.000
PFFC	3	5	99.061	0.020	4021	84.0	0.966	0.987	1.000
SIT	3	5	99.061	0.013	278	2.2	1.000	1.000	1.000
PF2	4	5	80.783	0.484	5767	47.0	0.892	0.945	0.999
PFFC	4	5	80.783	0.386	1402	32.0	0.946	0.978	1.000
SIT	4	5	80.783	0.262	160	1.4	0.997	0.991	1.000

Abbreviations: EFP, explained fit percentage; PF2, PARAFAC2–ALS; PFFC, PARAFAC2-flexible coupling; SIT, shift-invariant tri-linearity.

distinguishable for specific conditions (e.g., PF2 is slower on Conditions 1, 2, and 4 but significantly faster on Condition 3). The speed-up factor of SIT ranges between 19 and 28, comparing it to the fastest algorithm for each condition, respectively.

All algorithms provide very similar EFP values. The added noise puts an upper boundary on the EFP_{max} that can be achieved. Thus, the EFP values of the models (EFP_{model}) are evaluated relative to the EFP_{max} of a given dataset and noise level. If EFP_{model} (denoting the fit of one of the three algorithms on a specific dataset) exceeds EFP_{max}, the model is overfitted, and if EFP_{model} is below EFP_{max}, the model is underfitted. Therefore, the S/N ratio for each simulated dataset was calculated to determine EFP_{max}. Fit differences, EFP_{dif}, were calculated by subtracting EFP_{max} from EFP_{model}. The mean EFP_{max} values and the mean EFP_{dif} values are reported in Table 2. Negative EFP_{dif} values indicate a lack of fit, and positive values indicate overfitting of the models. From the results in Table 2, it becomes obvious that all models have a slight tendency to overfit the data. However, this tendency generally increases with higher noise levels and a larger number of components. For low noise levels, the EFP_{dif} values are undistinguishable between all models. For higher noise levels, PF2 shows the largest tendency for overfitting, followed by PFFC. The SIT algorithm has a lower tendency for overfitting compared with PFFC and PF2. It follows the expectations that PFFC and SIT have lower fit values than PF2 because they have been more constrained with non-negativity applied to all three modes instead of just two modes.

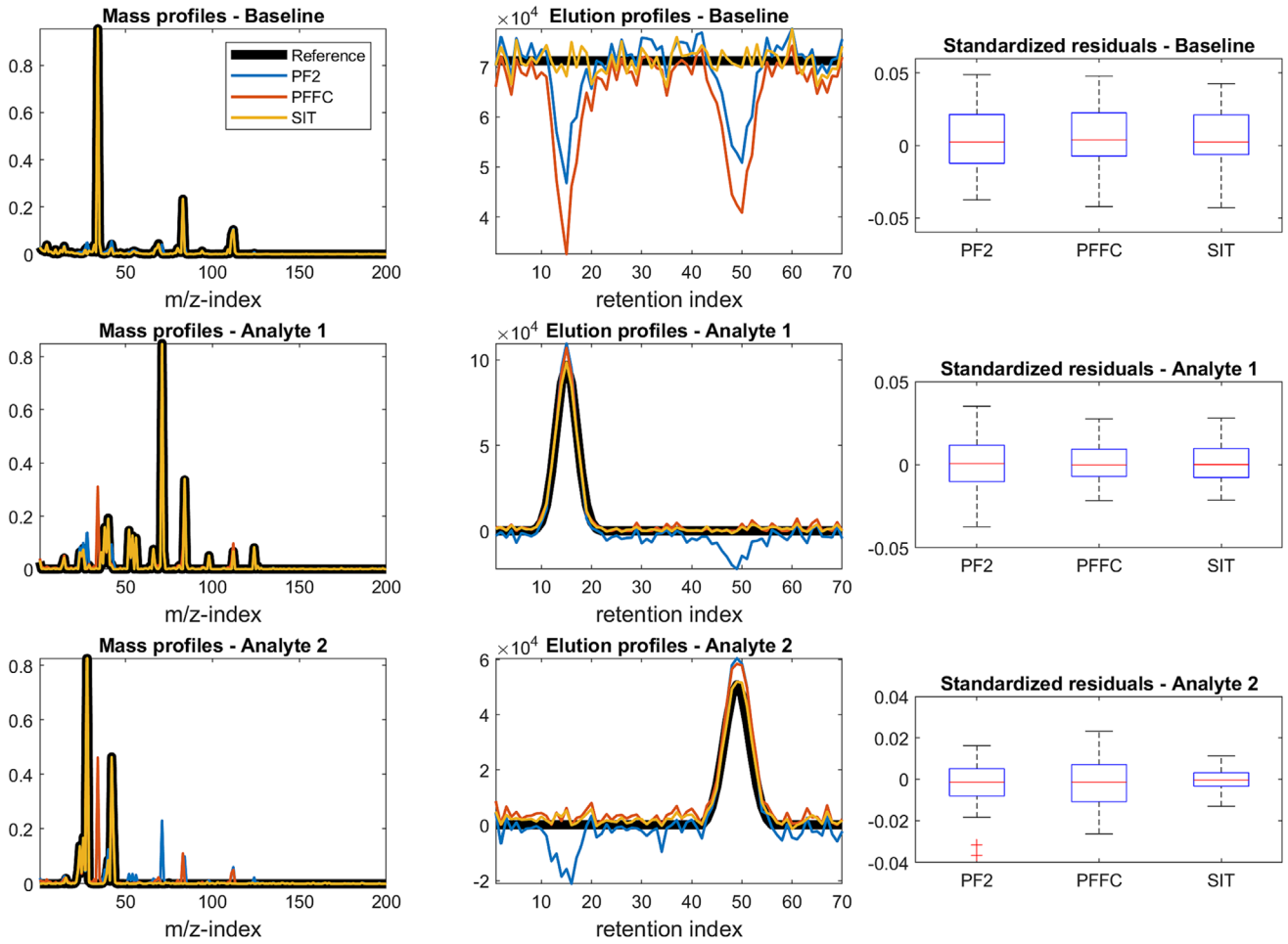


FIGURE 9 Randomly selected dataset to illustrate the differences in the estimated factors comparing PARAFAC2–ALS (PF2), PARAFAC2-flexible coupling (PFFC), and shift-invariant tri-linearity (SIT). The first two columns show overlays of the true factors (black) with the estimated factors from PF2 (blue), PFFC (red), and SIT (yellow). For better visualization, only one of the 50 elution profiles is plotted. The third column shows the distribution of the standardized residuals calculated by subtracting the true from the estimated concentrations.

The quality of the resolution was measured by the Tucker congruence[31] (cosine) between the modeled factors and the underlying true factors for each mode. Major differences in the quality of the resolution can be seen for modes **A** and **B** while all algorithms provide excellent estimates for mode **C**. All conditions considered, SIT provides the best resolution, while PF2 tends to perform better than PFFC for lower noise levels and vice versa for higher noise levels. The superior robustness of PFFC compared with PF2 for higher noise levels was also reported by Cohen et al.[12] However, except for the most challenging Condition 4, all algorithms provide good factor resolutions with Tucker congruences higher than 0.9.

In Figure 9, the true and estimated factors of a randomly picked dataset are shown. The first two columns of the matrix plot show the spectral profiles and the elution profiles. Deviations between true and modeled factors can be identified where the colored lines are not laying over the black line. The SIT estimates (yellow) resemble the true profiles very closely. Larger deviations can be seen for the PF2 (blue) and PFFC (orange) estimates. The artifactual peaks in the mass profile estimates of PF2 and PFFC appear at m/z values, at which the baseline signal has strong bands. At the same time, the estimates for the elution profiles of the baseline show "bumps" at the positions of the analyte peaks. The mass profile estimates from PF2 confuse some of the bands between analytes 1 and 2, which is also reflected in the elution profiles showing negative peaks. In summary, it appears that SIT has less difficulties in unmixing the signal contributions from the true underlying factors than PF2 and PFFC have. The third column shows the distributions of the standardized concentration residuals. The residuals have been calculated by subtracting the true concentrations from the estimated concentrations and scaling the result by the true concentration. The quantitative accuracy of the different algorithms is comparable, with the exception that SIT yields higher accuracy on analyte 2.

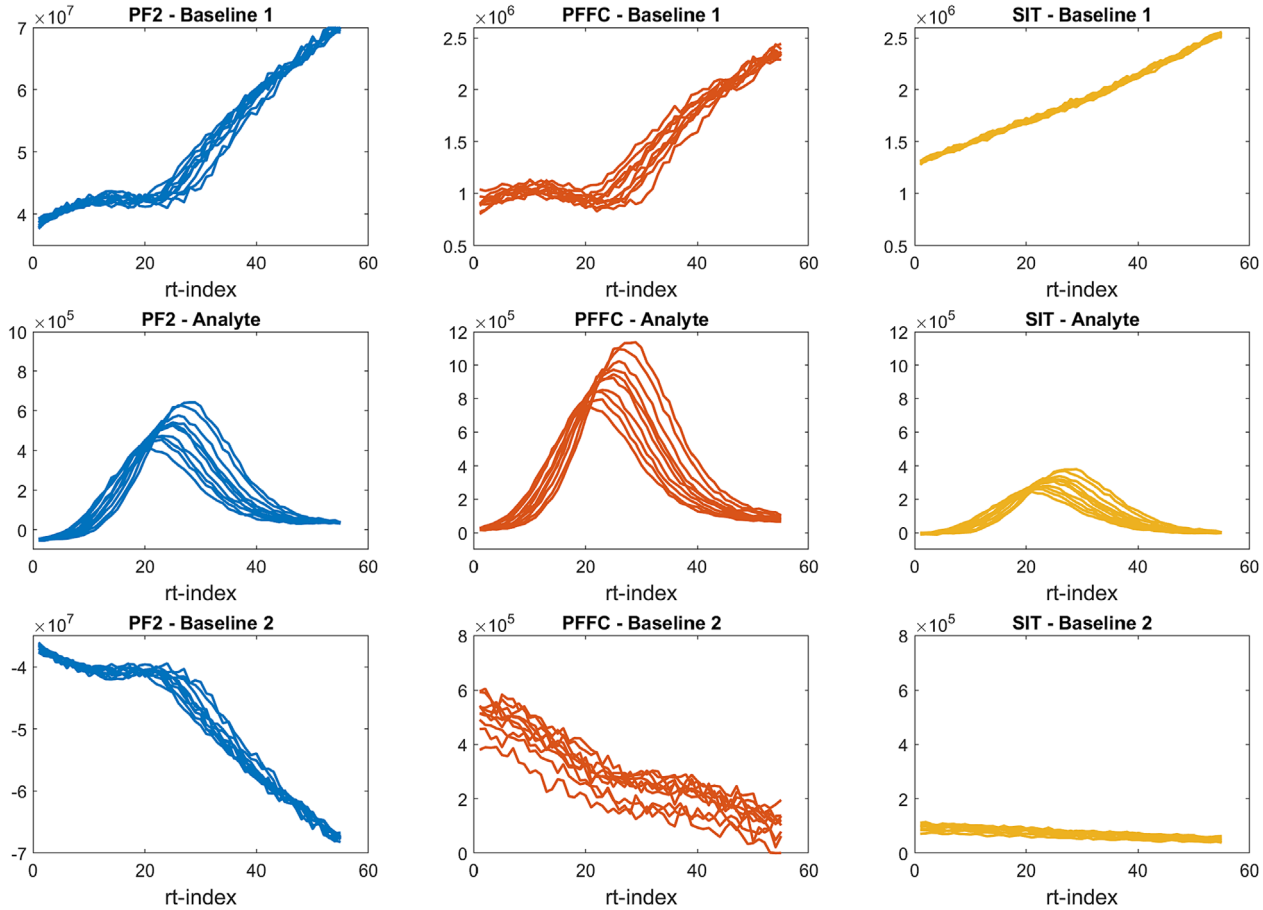


FIGURE 10 Selected elution profiles of different three-component models fitted to Interval 12 of the apple wine dataset (cf. Figure 8). The upper and lower rows show baseline components, and the middle row shows the analyte signal.

4.2 | Apple wine

In order to challenge the results of the simulation study, SIT was benchmarked against PF2 and PFFC on three different real GC-MS datasets. On average, SIT was 65 times faster than PF2 and 58 times faster than PFFC. A detailed summary of the computation times for all three datasets is given in Appendix B.

The first dataset (left-most in Figure 8) is a small peak on a baseline with a slope. The dataset was fitted with three component models: PF2, PFFC, and SIT. The residuals for this model indicate that all chemical information has been captured. Because there is no reference available, the modeled profiles are compared qualitatively. In Figure 10, selected elution profiles are illustrated for the three fitted components, and in Figure 11, the respective mass spectra are shown. The results indicate that PF2 is suffering from two-factor degeneracy (2-FD), as the elution profiles for the baselines are highly negative and the corresponding mass spectra are highly positively correlated.[32] The imposed non-negativity constraints on the elution profiles save the solutions of PFFC and SIT from becoming degenerate.[33] Nevertheless, PFFC has difficulties unmixing the contributions of baseline and analyte signals. This can be emphasized by the "bumps" in the elution profiles of baseline 1. Moreover, the mass signals of baseline 1 are also very pronounced in the mass spectra of the analyte.

The analyte was identified as butyl acetate by a NIST library search. The Tucker congruences of the mass profiles with the normalized reference spectra are 0.519, 0.279, and 0.985 for PF2, PFFC, and SIT, respectively. Interestingly, the Tucker congruence between modeled and reference spectra does not change going from a three-component SIT model to a two-component SIT model. Moreover, the mass profiles of a four-component PFFC model have a drastically higher Tucker congruence with the butyl acetate reference spectrum (0.969).

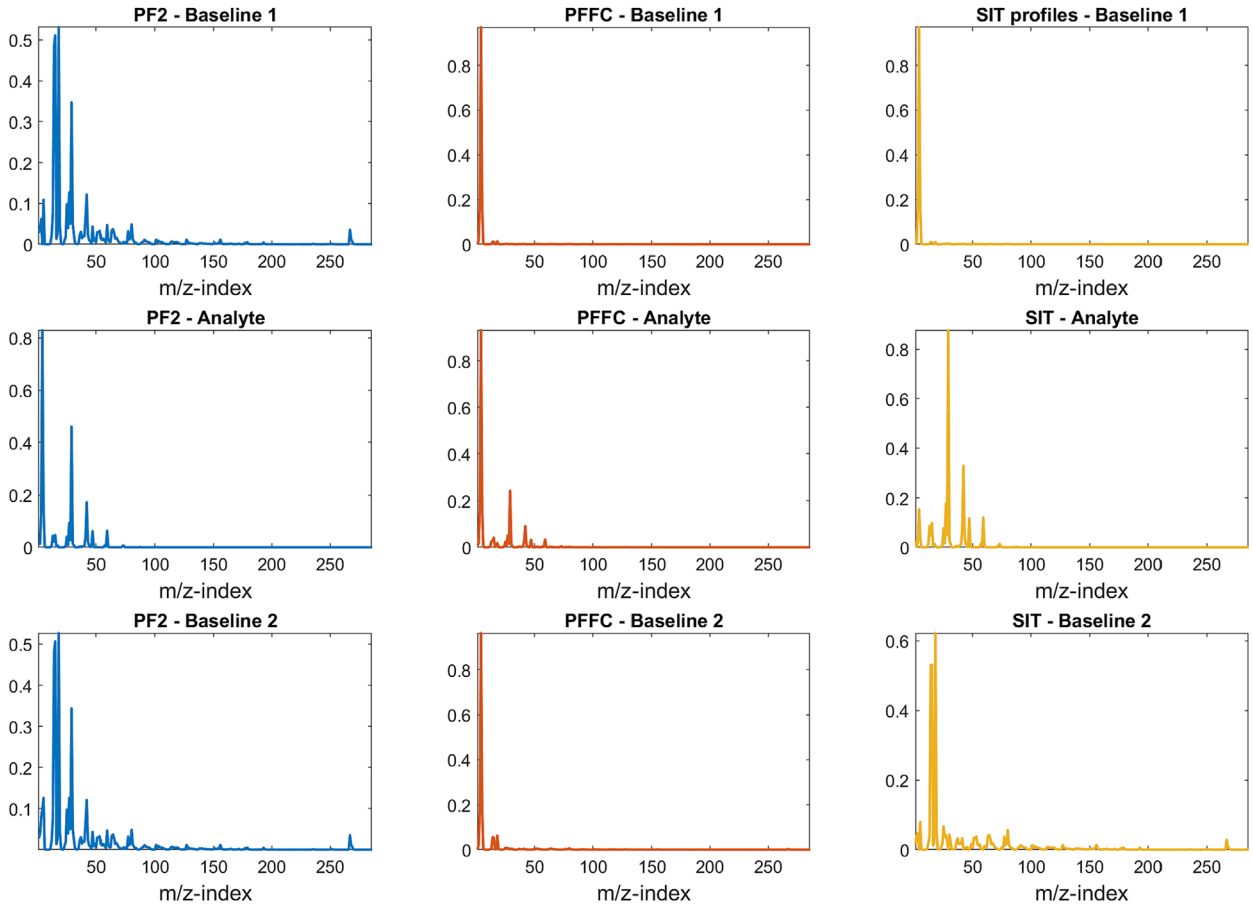


FIGURE 11 Mass spectra belonging to the elution profiles shown in Figure 10. The upper and lower rows show the mass spectra of baseline components, and the middle row shows the analyte signal. The analyte mass spectra of PARAFAC2–ALS (PF2) and PARAFAC2-flexible coupling (PFFC) both have a strong band at low m/z-index, which also appears in the baseline mass spectra. This is not the case for the analyte signal of shift-invariant tri-linearity (SIT).

The second dataset (middle in Figure 8) shows an intense, heavily shifted peak signal. A three-component model was selected as being optimal because models with more components resulted in the elution profiles being distorted. In Figure 12, selected elution profiles of the three analyte signals are shown. The PF2 model suffers, as in the previous example, from 2-FD, as can be seen from the elution profiles of analytes 2 and 3. The results from PFFC and SIT look very similar, especially considering the modeled mass profiles shown in Figure 13. However, the elution profiles of analyte 2 modeled with PFFC are distorted. These artifacts get worse with increasing intensities of the analyte 3 signal. Another observation is that the peak shapes of analyte 3 are clearly varying with increasing intensity, from an almost Gaussian shape (low intensity) to a triangular peak shape (high intensity). As outlined in Section 2.4, SIT can basically not model shape changes. If shape changes are present, the power spectra matrix $\widehat{\mathbf{M}}$ will have a rank greater than one. Hence, approximating $\widehat{\mathbf{M}}$ by a one-component PCA model results in truncating the Fourier coefficients. The truncation can cause artifacts in the data such as ringing, for example, related to Gibbs phenomena. Considering this limitation, it is quite surprising how well SIT resolves the elution profiles. Ringing, according to Gibbs phenomena, is present in the elution profiles of analytes 1 and 3, but only gets visible in Figure 12 after zooming in. However, PFFC and SIT fail to model the baseline properly, as can be seen from the offset present in the elution profile of analyte 2.

One limitation of the SIT model becomes obvious when looking at the results achieved by fitting a seven-component model to the third dataset (right in Figure 8). The large cut-off peak causes a discontinuity at the edge of the interval, which cannot be modeled properly by the DFT. The reason is that the discrete time signal is required to be periodic. This means that we must have a smooth transition if we would connect the starting point of the interval with the endpoint. Therefore, we obtain large residuals at the edges of the interval when we model the third dataset with the SIT

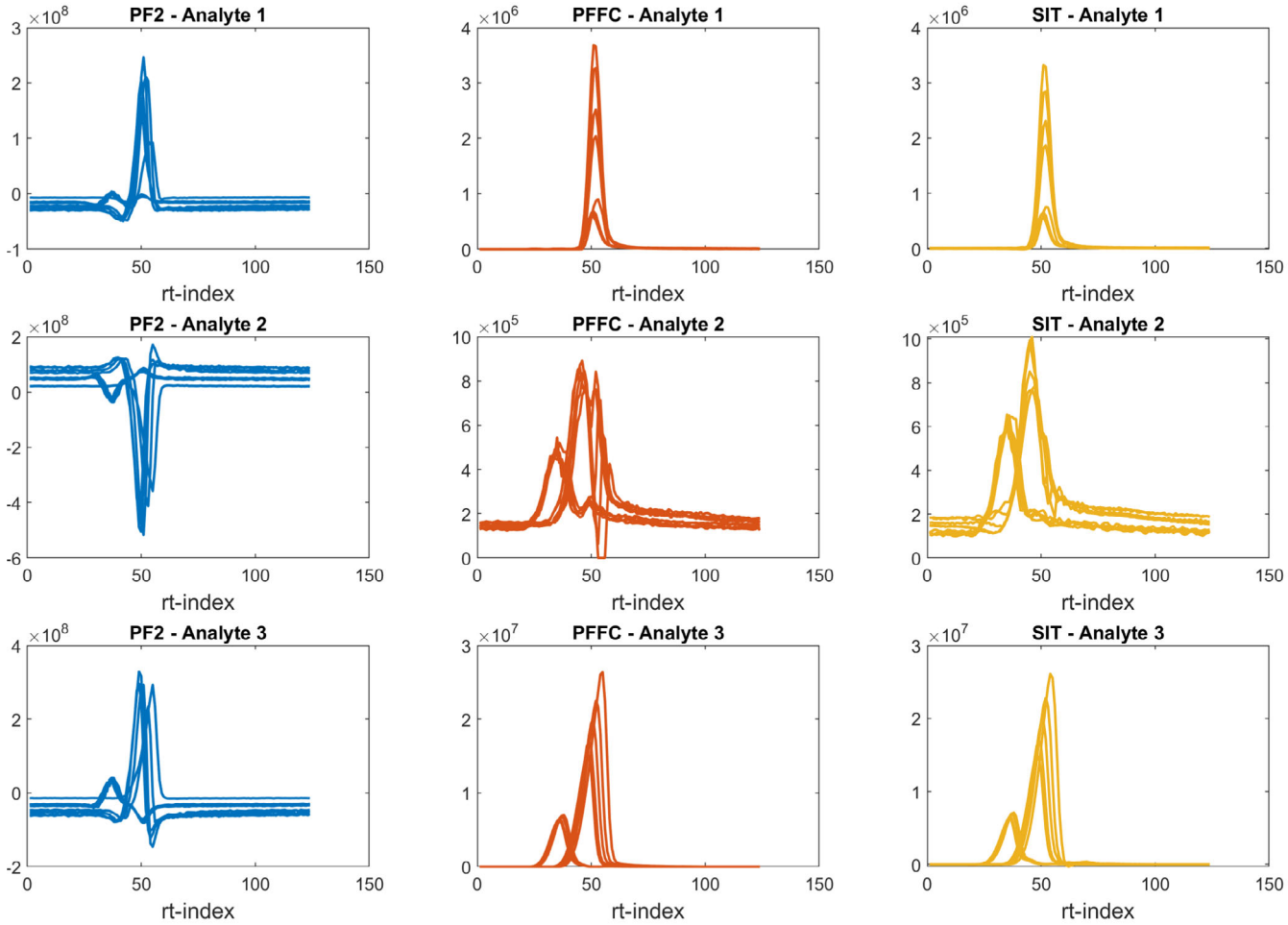


FIGURE 12 Selected elution profiles of different three-component models fitted to Interval 18 of the apple wine dataset (cf. Figure 8). The factors modeled with PARAFAC2–ALS (PF2) suffer from two-factor degeneracy (2-FD). The results of PARAFAC2-flexible coupling (PFFC) and shift-invariant tri-linearity (SIT) are overall similar. The elution profiles modeled with PFFC for analyte 2 show artifacts.

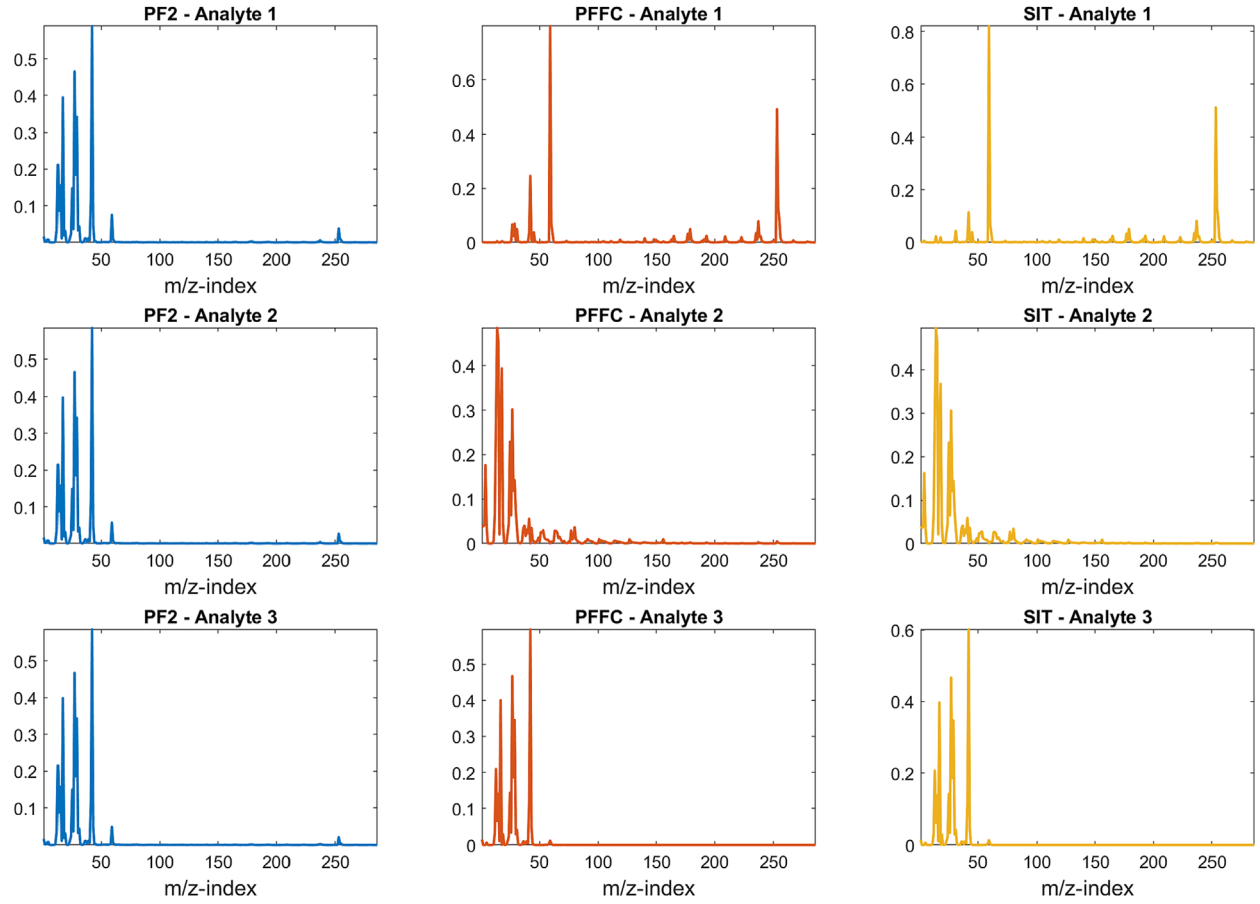


FIGURE 13 Mass spectra belonging to the elution profiles shown in Figure 12. The mass spectra of PARAFAC2-flexible coupling (PFFC) and shift-invariant tri-linearity (SIT) compare very well across all analytes. The mass spectra obtained with PARAFAC2–ALS (PF2) are all the same, indicating factor degeneracy.

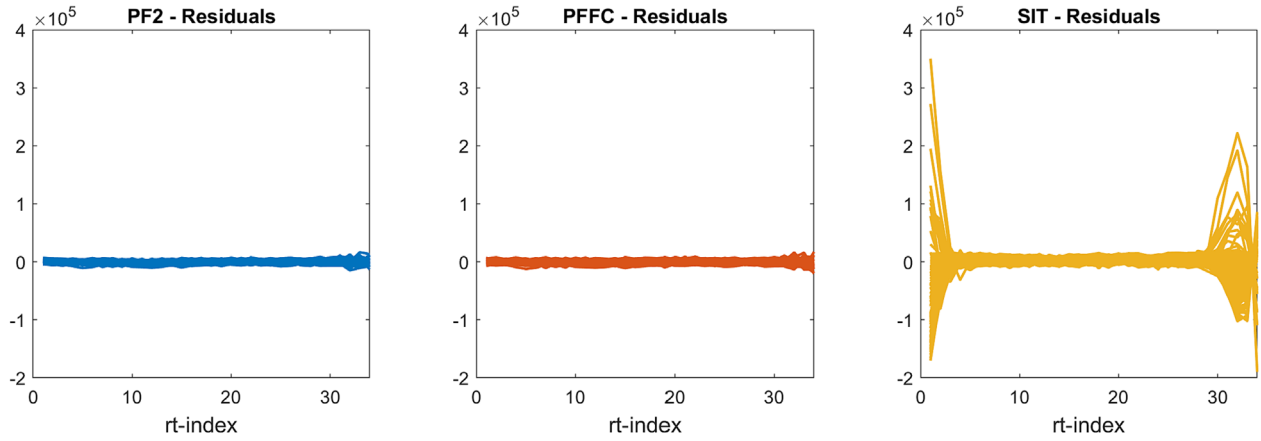


FIGURE 14 Residuals of different models fitted to Interval 33 (cf. Figure 8). The model residuals for shift-invariant tri-linearity (SIT) are significantly larger than the residuals obtained with PARAFAC2–ALS (PF2) and PARAFAC2-flexible coupling (PFFC).

algorithm (cf. Figure 14). Further investigation revealed that the large residuals are exclusively caused by one factor describing the large cut-off peak, while the other factors do not show any artifacts. Thus, SIT can still model the elution profiles of the analytes hidden in the tail of the large peak very well, as can be seen in Figure 15. The mass spectra of analytes 1 and 2 were compared against the NIST library. No reliable match could be found for analyte 1. The best

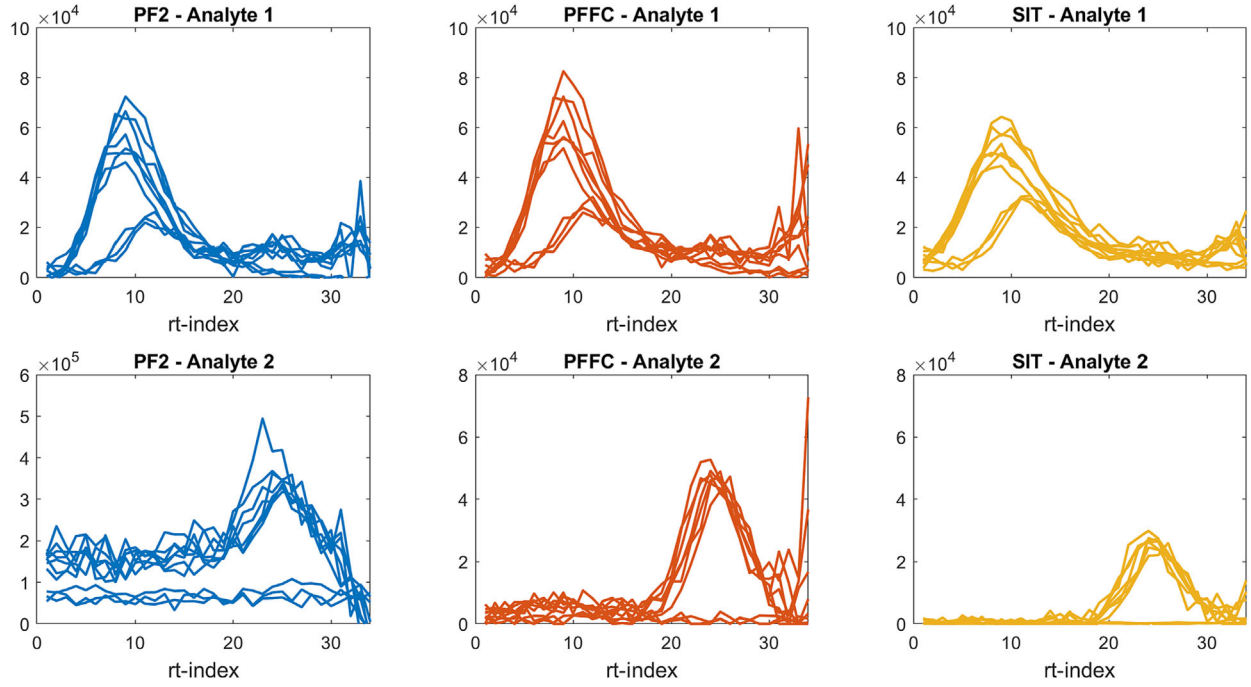


FIGURE 15 Selected elution profiles of different three-component models fitted to Interval 33 of the apple wine dataset (cf. Figure 8). All models resolve very similar elution profiles for analyte 1. The elution profiles of analyte 2 compare well between PARAFAC2-flexible coupling (PFFC) and shift-invariant tri-linearity (SIT). The elution profiles obtained with PARAFAC2–ALS (PF2) for analyte 2 look distorted and have ten times larger weights than the elution profiles from PFFC and SIT.

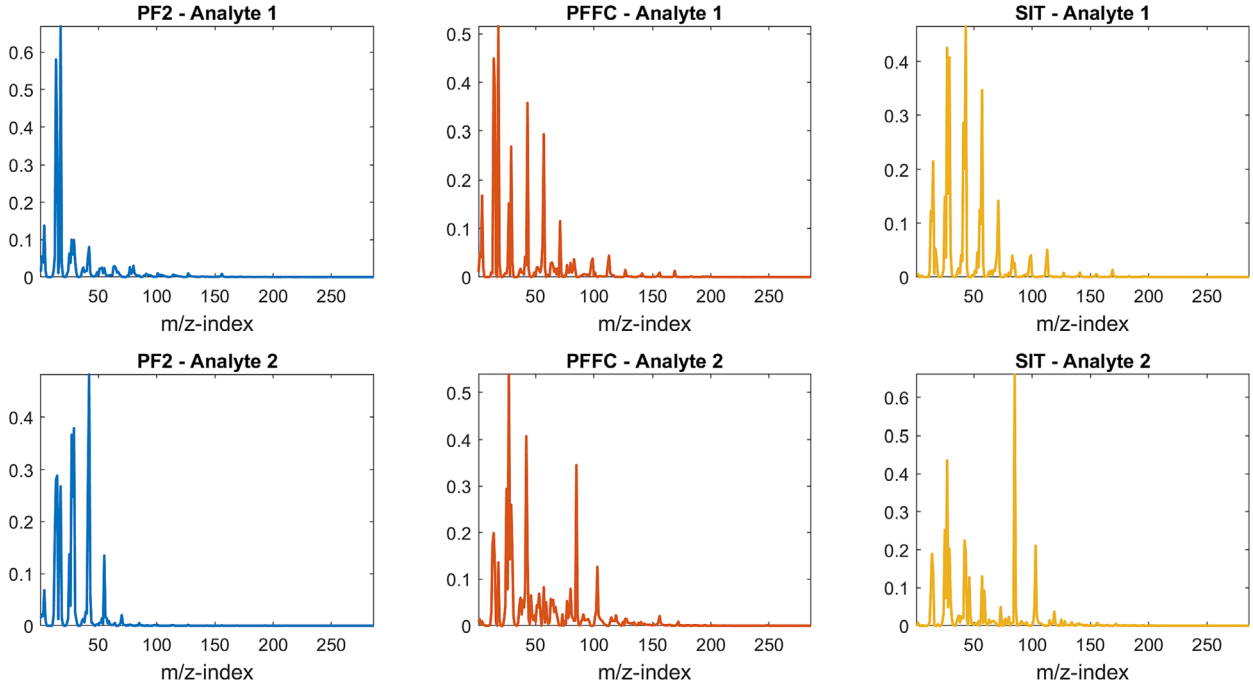


FIGURE 16 Mass spectra belonging to the elution profiles shown in Figure 15. The mass spectra of PARAFAC2-flexible coupling (PFFC) and shift-invariant tri-linearity (SIT) compare very well across all analytes. The mass spectra obtained with PARAFAC2–ALS (PF2) are distinct from the solutions obtained with PFFC and SIT.

match for analyte 2 was found to be hexanoic acid butyl ester for SIT (Tucker congruence 0.88) and hexanoic acid 2-methylpropyl ester for PFFC (Tucker congruence 0.80). Because the PF2 solution suffered from 2-FD and the mass spectra looked very unlike the mass spectra of PFFC and SIT, no plausible match was found in either case (cf. Figure 16).

5 | CONCLUSION

This paper demonstrated that SIT provides a fast, accurate alternative to existing PARAFAC2 algorithms[10,12–14,18] for untargeted modeling of hyphenated chromatography data. In contrast to classical PARAFAC2-ALS,[10] SIT allows constraints in the shifted mode, for example, non-negativity. Compared with existing SIT methods,[18] the proposed SIT approach models shifted data implicitly. The quality of resolution of the proposed SIT method was evaluated on simulated[17] and real GC-MS datasets.[11] Compared with PARAFAC2-ALS, the SIT method performed better on simulated and real GC-MS datasets. On challenging real GC-MS datasets, non-negativity constraints on the shifted mode were necessary to prevent degenerate solutions. In comparison with PFFC,[12] SIT performed better on the simulated datasets and, in some cases, better on real GC-MS data. Specifically, SIT achieved better factor resolution in cases with large baseline components. Moreover, SIT was more efficient on some datasets because fewer components were required to capture the chemically meaningful information. However, other than peak shifts, the currently proposed SIT algorithm does not account for deviations from tri-linearity, such as significant shape changes in the elution profiles. Although SIT handled shape changes surprisingly well (cf. Figures 12 and 13), further improvements to the algorithm can certainly extend the application range of the algorithm further. The SIT algorithm in its current form has been implemented in the PARADISe toolbox (https://ucphchemometrics.com/paradise, [December 15, 2022]).

NOTATION

x_{ijk}	scalar,
$\mathbf{x}$	$(I \times 1)$ vector,
$\mathbf{x}^{\mathrm{T}}$	$(1 \times I)$ transpose of $\mathbf{x}$,
$\mathbf{X}$	$(I \times J)$ matrix,
$\underline{\mathbf{X}}$	$(I \times J \times K)$ tensor,
$\odot$	Kathri–Rao product (column-wise Kronecker product),
$\circ$	element-wise matrix multiplication (Hadamard product),
$\|\cdot\|_F^2$	Frobenius norm,
$\widehat{f}(k)$	Fourier transform of $f(n)$,
$\widehat{f}^{*}(k)$	complex conjugate of $\widehat{f}(k)$

DATA AVAILABILITY STATEMENT

Data and algorithms will be freely available at https://ucphchemometrics.com/.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3501.

ORCID

Paul-Albert Schneide https://orcid.org/0000-0002-8365-5699
Rasmus Bro https://orcid.org/0000-0002-7641-4854

REFERENCES

1. Harshman, R. A. *Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-model Factor Analysis*. in (1970).
2. Carroll JD, Chang J-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*. 1970;35(3):283-319. doi:10.1007/BF02310791
3. Sidiropoulos ND, Bro R, Giannakis GB. Parallel factor analysis in sensor array processing. *IEEE Trans Signal Process*. 2000;48(8):2377-2388. doi:10.1109/78.852018

4. Harshman RA, Lundy ME. The PARAFAC model for three-way factor analysis and multidimensional scaling. In: Law HG, Snyder CW Jr, Hattie JA, McDonald RP, eds. *Research Methods for Multimode Data Analysis*. Praeger; 1984:122-215.
5. Kruskal, J. Rank B., *Decomposition, and Uniqueness for 3-way and N-way Arrays*. in (1989).
6. Bro R. PARAFAC. Tutorial and applications. *Chemom Intel Lab Syst*. 1997;38(2):149-171. doi:10.1016/S0169-7439(97)00032-4
7. Bro R, Andersson CA, Kiers HAL. PARAFAC2—part II. Modeling chromatographic data with retention time shifts. *J Chemometr*. 1999; 13(3-4):295-309. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4%3C295::AID-CEM547%3E3.0.CO;2-Y
8. Skov T, Bro R. Solving fundamental problems in chromatographic analysis. *Anal Bioanal Chem*. 2008;390(1):281-285. doi:10.1007/s00216-007-1618-z
9. Harshman RA. *PARAFAC2: Mathematical and technical notes*. UCLA Working Papers in Phonetics. Vol. 22. University Microfilms; 1972:30-44. No. 10,085.
10. Kiers HAL, ten Berge JMF, Bro R. PARAFAC2—part I. A direct fitting algorithm for the PARAFAC2 model. *J Chemometr*. 1999;13(3-4): 275-294. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4%3C275::AID-CEM543%3E3.0.CO;2-B
11. Baccolo G, Quintanilla-Casas B, Vichi S, Augustijn D, Bro R. From untargeted chemical profiling to peak tables—a fully automated AI driven approach to untargeted GC-MS. *TrAC Trends Anal Chem*. 2021;145:116451. doi:10.1016/j.trac.2021.116451
12. Cohen JE, Bro R. Nonnegative PARAFAC2: a flexible coupling approach. In: Yannick D, Gannot S, R M, D. PM, D. W, eds. *Latent Variable Analysis and Signal Separation*. Springer International Publishing; 2018:89-98. doi:10.1007/978-3-319-93764-9_9
13. Van Benthem MH, Keller TJ, Gillispie GD, DeJong SA. Getting to the core of PARAFAC2, a nonnegative approach. *Chemom Intel Lab Syst*. 2020;206:104127. doi:10.1016/j.chemolab.2020.104127
14. Roald M, Schenker C, Calhoun VD, et al. An AO-ADMM approach to constraining PARAFAC2 on all modes. *SIAM J Math Data Sci*. 2022;4(3):1191-1222. doi:10.1137/21M1450033
15. Yu H, Bro R, Gallagher NB. PARASIAS: a new method for analyzing higher-order tensors with shifting profiles. *Anal Chim Acta*. 2023; 1238:339848. doi:10.1016/j.aca.2022.339848
16. Yu H, Augustijn D, Bro R. Accelerating PARAFAC2 algorithms for non-negative complex tensor decomposition. *Chemom Intel Lab Syst*. 2021;214:104312. doi:10.1016/j.chemolab.2021.104312
17. Tian K, Wu L, Min S, Bro R. Geometric search: a new approach for fitting PARAFAC2 models on GC-MS data. *Talanta*. 2018;185:378-386. doi:10.1016/j.talanta.2018.03.088
18. Zhang X, Tauler R. Flexible implementation of the trilinearity constraint in multivariate curve resolution alternating least squares (MCR-ALS) of chromatographic and other type of data. *Molecules*. 2022;27(7):2338. doi:10.3390/molecules27072338
19. Tomasi G, Bro R. A comparison of algorithms for fitting the PARAFAC model. *Comput Stat Data Anal*. 2006;50(7):1700-1734. doi:10.1016/j.csda.2004.11.013
20. Tauler R, Marqués I, Casassas E. Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths. *J Chemometr*. 1998;12(1):55-75. doi:10.1002/(SICI)1099-128X(199801/02)12:1%3C55::AID-CEM501%3E3.0.CO;2-#
21. Tauler R, Smilde A, Kowalski B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J Chemometr*. 1995;9(1):31-58. doi:10.1002/cem.1180090105
22. de Juan A, Tauler R. Multivariate curve resolution: 50 years addressing the mixture analysis problem—a review. *Anal Chim Acta*. 2021; 1145:59-78. doi:10.1016/j.aca.2020.10.051
23. de Juan A, Tauler R. Comparison of three-way resolution methods for non-trilinear chemical data sets. *J Chemometr*. 2001;15(10):749-771. doi:10.1002/cem.662
24. Tauler R. Multivariate curve resolution of multiway data using the multilinearity constraint. *J Chemometr*. 2021;35:e3279. doi:10.1002/cem.3279
25. Cooley JW, Tukey JW. An algorithm for the machine calculation of complex Fourier series. *Math Comput*. 1965;19(90):297-301. doi:10.1090/S0025-5718-1965-0178586-1
26. Brunton S, Kutz N. Fourier and wavelet transforms. In: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press; 2019:47-83. doi:10.1017/9781108380690.003
27. Naish PJ, Hartwell S. Exponentially modified Gaussian functions—a good model for chromatographic peaks in isocratic HPLC? *Chromatographia*. 1988;26(1):285-296. doi:10.1007/BF02268168
28. Tavakkoli E, Abdollahi H, Gemperline PJ. Soft-trilinear constraints for improved quantitation in multivariate curve resolution. *Analyst*. 2020;145(1):223-232. doi:10.1039/C8AN00615F
29. Johnsen LG, Skou PB, Khakimov B, Bro R. Gas chromatography–mass spectrometry data processing made easy. *J Chromatogr a*. 2017; 1503:57-64. doi:10.1016/j.chroma.2017.04.052
30. Yu H, Bro R. PARAFAC2 and local minima. *Chemom Intel Lab Syst*. 2021;219:104446. doi:10.1016/j.chemolab.2021.104446
31. Tucker, L. R. *A Method for Synthesis of Factor Analysis Studies*. https://apps.dtic.mil/sti/citations/AD0047524 (1951).
32. Mitchell BC, Burdick DS. Slowly converging PARAFAC sequences: swamps and two-factor degeneracies. *J Chemometr*. 1994;8(2):155-168. doi:10.1002/cem.1180080207
33. Lim, L.-H. *Optimal Solutions to Non-negative PARAFAC/Multilinear NMF Always Exist*. (2005).

How to cite this article: Schneide P-A, Bro R, Gallagher NB. Shift-invariant tri-linearity—A new model for resolving untargeted gas chromatography coupled mass spectrometry data. *Journal of Chemometrics*. 2023;37(8): e3501. doi:10.1002/cem.3501

APPENDIX A: Detailed description of the SIT algorithm in MATLAB pseudocode

Algorithm: Shift-invariant tri-linearity

0. Initialize $\mathbf{V}$ and SSE, e.g.:
   $\mathbf{V} = \mathrm{rand}(J, K\,x\,R)$
   $SSE = 2\|\underline{\mathbf{X}}\|_F^2$
1. Update $\mathbf{S}^{\mathrm{T}}$
   $\mathbf{S}^{\mathrm{T}} = pinv\left(\mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\mathbf{V}^{\mathrm{T}}\mathbf{X}_{BC}$
   $for\ r = 1:R$
   $\mathbf{S}^{\mathrm{T}}(r,:) = \frac{\mathbf{S}^{\mathrm{T}}(r,:)}{\left\|\mathbf{S}^{\mathrm{T}}(r,:)\right\|_F^2}$
   end
2. Update $\mathbf{V}$

   2a. Least Squares Estimate
   $\mathbf{V} = \mathbf{X}_{BC}^{\mathrm{T}}\mathbf{S}\,pinv\left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)$
   2b. Applying Shift-Invariant-Tri-Linearity Constraint
   $for\ r = 1:R$
   $\mathbf{v}_r = \mathbf{V}(:,r)$
   $\mathbf{M} = reshape(\mathbf{v}_r, [J,K])^{\mathrm{T}}$
   $\mathbf{Z} = fft(\mathbf{M})$
   $\widehat{\mathbf{M}} = abs(\mathbf{Z})$
   $\boldsymbol{\phi} = \mathrm{angle}\left(\widehat{\mathbf{M}}\right)$
   $[\mathbf{U},\mathbf{T},\mathbf{W}] = SVD\left(\widehat{\mathbf{M}}\right)$
   $\widehat{\mathbf{M}}_{\mathrm{est}} = \mathbf{U}(:,1)\mathbf{T}(1,1)\mathbf{W}(:,1)^{\mathrm{T}}$
   $\mathbf{M}_{\mathrm{est}} = real\left(ifft\left(\widehat{\mathbf{M}}_{\mathrm{est}} \circ \boldsymbol{\phi}\right)\right)$
   $\mathbf{M}_{\mathrm{est}} = \mathbf{M}_{\mathrm{est}}^{\mathrm{T}}$
   $\mathbf{V}(:,r) = \mathbf{M}_{\mathrm{est}}(:)$
   end

3. Calculate $\mathbf{X}_{\mathrm{est}}$ and SSE_{new}
   $\mathbf{X}_{\mathrm{est}} = \mathbf{V}\mathbf{S}^{\mathrm{T}}$
   $SSE_{\mathrm{new}} = \left\|\mathbf{X}_{\mathrm{est}} - \mathbf{X}_{BC}^{\mathrm{T}}\right\|_F^2$
4. Evaluate loss function
   if $ConvCrit > SSE - SSE_{\mathrm{new}}$ → **Stop**

   else $SSE = SSE_{new}$ → return to **1.**

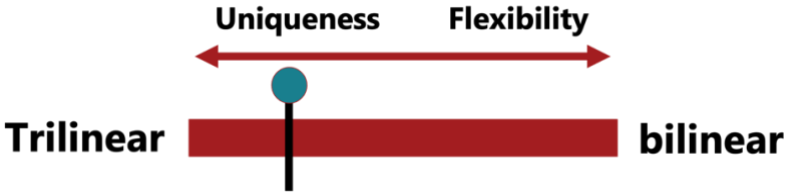
APPENDIX B: Summary of computation times for the apple wine datasets

TABLE 3 Summary shows the array sizes of the respective intervals, the number of components that have been selected as well as the mean computation time and standard deviations considering 10 repetitive fits starting from random initial values.

	Array size	# comps.	CT PF2 (s)	CT PFFC (s)	CT SIT (s)
Interval 12	$155 \times 55 \times 286$	3	199 ± 32	217 ± 82	3.1 ± 0.6
Interval 18	$155 \times 124 \times 286$	3	249 ± 1.3	44 ± 2	4.8 ± 1.7
Interval 33	$155 \times 34 \times 286$	7	428 ± 112	504 ± 193	5.3 ± 1.4

Paper 2

Schneide P-A, Gallagher NB, Bro R. Shift invariant soft trilinearity: Modelling shifts and shape changes in gas-chromatography coupled mass spectrometry. Chemometrics and Intelligent Laboratory Systems. 2024; 251. doi: 10.1016/j.chemolab.2024.105155.

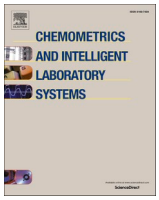


Chemometrics and Intelligent Laboratory Systems 251 (2024) 105155

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Shift invariant soft trilinearity: Modelling shifts and shape changes in gas-chromatography coupled mass spectrometry

Paul-Albert Schneide [a,b,*], Neal B. Gallagher [c], Rasmus Bro [a,**]

[a] Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958, Frederiksberg C, Denmark
[b] Analytical Science Department, BASF SE, Carl-Bosch-Strasse 38, 67056, Ludwigshafen am Rhein, Germany
[c] Eigenvector Research Inc., Manson, WA, 98831, USA

1. Introduction

Bilinear, trilinear and multilinear methods have been extensively discussed and used for the analysis of hyphenated chromatographic data [1–9]. Data sets can be acquired on analytical platforms such as gas-chromatography coupled mass spectrometry, liquid chromatography with diode-array detection or liquid chromatography coupled mass spectrometry. The aim of applying chemometrics methods is to facilitate the analysis of these highly complex data sets. More specifically, to retrieve the maximum amount of information as accurately and as fast as possible. Chemometric methods therefore have to be able to deconvolute peaks and extract the qualitative (analyte spectra) and quantitative (peak areas) information.

The deconvolution procedure comprises separation of overlapped peaks from each other and from the baseline signal. Difficulties in the signal separation arise from the imperfections of the measurement process. For instance, elution profiles of analytes may change positions and shapes in chromatographic measurements across multiple runs. Consequently, data structures often deviate from perfect trilinearity limiting the utility of trilinear methods. However, methods like PARAllel FACtor analysis 2 (PARAFAC2) and shift-invariant versions of trilinearity constrained MCR models can accommodate deviations from a trilinear data structure caused by shifts in the elution profiles [8,10,11]. Additional deviations from trilinearity due to shape changes can be modeled by more flexible PARAFAC2 implementations and with bilinear models [12,13]. One problem related to bilinear models is the rotational ambiguity which may cause the resolved factors to be less accurate, unless additional constraints are applied to narrow the solution space (area of feasible solutions) [14–17]. In mixed bilinear-trilinear models, it is possible to constrain the different factors separately [10,13]. However, modelling a factor with a bilinear or trilinear model is a rather black or white decision. The model complexity best describing real data most certainly lies somewhere between these poles. Consequently, the selected model may either be too flexible or too rigid.

With this work we propose a new shift-invariant soft-trilinearity constraint that automatically adjusts the flexibility of the model on-the-fly. Using real thermo-desorption (TD) GC-MS data, we demonstrate that our approach successfully recovers elution profiles of overlapped peaks that exhibit shifts and shape changes across measurements. Moreover, we show that our method suffers less from rotational ambiguity and local minima than non-negativity constraint MCR and PARAFAC 2 flexible coupling, despite its flexibility.

2. Background

A set of ideal GC-MS samples follows a trilinear data structure. However, because the chromatographic conditions cannot be kept perfectly constant across multiple measurements the retention times will vary across a set of samples. Further, non-linearities in the partition isotherms can cause deviations from ideal Gaussian peak shapes. The shape of the resulting peaks in the non-linear regime of the partition isotherms depends on the concentration and the physico-chemical properties of the analyte, the properties of the stationary phase and the underlying support material (beside other factors) [18]. Unfortunately, modelling data which does not follow a trilinear structure with a trilinear model will deliver inaccurate results. This is exemplified in Fig. 1, which shows the trilinearity constrained MCR model applied to shifted-trilinear data. The trilinearity constrained MCR model is equivalent to the PARAFAC model but has the advantage that the trilinearity constraint can be applied component-wise. In Fig. 1, the trilinearity constraint is applied to the blue colored component. It is required that the reshaped elution profile can be approximated well by a one-component PCA model for the constraint to hold. This is certainly not the case, since at least two components are required to model the data: one component that describes the mean shape of the elution profiles and one component that models the shift as shown in the bottom of

* Corresponding author. Analytical Science Department, BASF SE, Carl-Bosch-Strasse 38, 67056, Ludwigshafen am Rhein, Germany.
** Corresponding author. Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958, Frederiksberg C, Denmark.
E-mail addresses: paul.schneide@basf.com (P.-A. Schneide), rb@food.ku.dk (R. Bro).

https://doi.org/10.1016/j.chemolab.2024.105155
Received 20 December 2023; Received in revised form 30 May 2024; Accepted 2 June 2024
Available online 8 June 2024

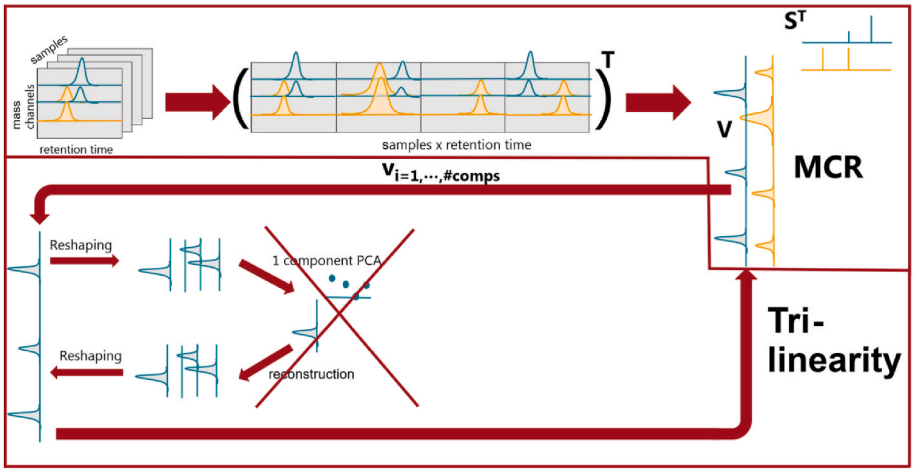


Fig. 1. Visualization of a trilinearity constrained MCR model. The constraint is violated because of the retention time shifts that occur between different measurements.

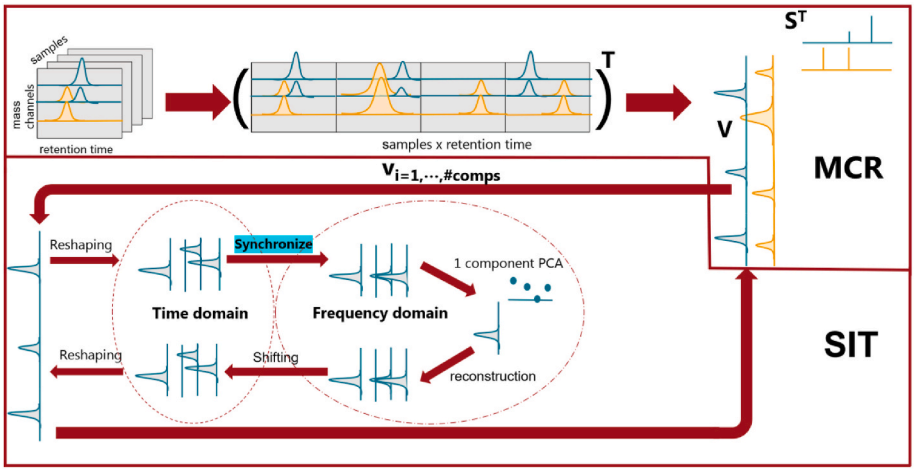


Fig. 2. The shift-invariant trilinearity constraint provides accurate and unique factor estimates for shifted-trilinear data. The synchronization of the elution profiles is achieved via domain transfer from time to frequency domain.

Fig. 1.

The PARAFAC2 model provides an elegant but non-trivial approach to model specific kinds of shifted-trilinear data. The trick is, to require constant cross-products of the elution profiles rather than constant peak positions across samples [1,11,19]. The drawback of this approach is its algorithmic complexity and the fact that only linear shifts of the elution profiles that do not violate the cross-product-constraint are allowed [20–22]. Nevertheless, the usefulness of this approach has been proven in many applications [23–26]. Another way of removing the shift in the MCR trilinearity framework is, to synchronize the elution profiles prior to performing the PCA. The synchronization can either be performed by aligning the elution profiles or by doing a domain transfer from time to frequency domain using the discrete Fourier Transform [10,27]. The latter method exploits the shift-invariant property of the amplitude spectrum [28]. Hence, the shift-invariant amplitude spectra of the elution profiles shown in Fig. 1 can be perfectly reconstructed by a one component PCA model, as shown in Fig. 2. The original, shifted spectra are reconstructed by applying the inverse Discrete Fourier Transform on the one-component estimate of the amplitude spectra and the phase spectra. This method is described in a simplified scheme (neglecting the handling of the phase spectra) in Fig. 2. This is implemented in the shift-invariant trilinearity (SIT) approach, that has been described as computationally more efficient and sometimes more parsimonious than the PARAFAC2 model regarding the number of latent variables [8]. It is

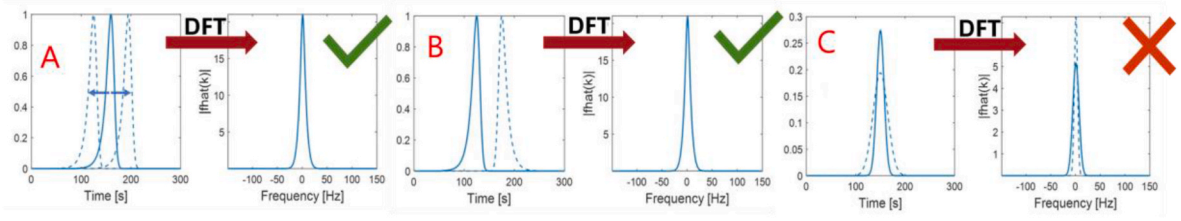


Fig. 3. Modifications of a time domain signal are compared with the resulting changes in the amplitude spectrum. Modifications A) and B) do not change the amplitude spectrum whereas modification C) results in a change of the amplitude spectrum.

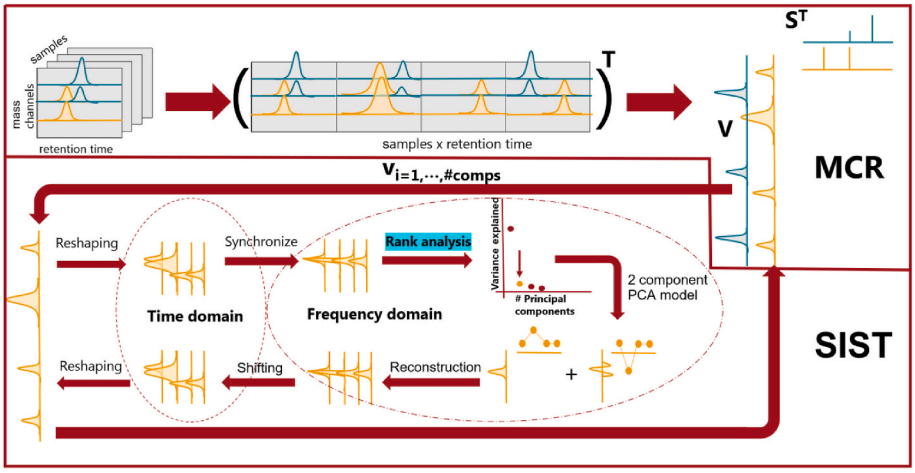


Fig. 4. Visualization of the shift-invariant soft-trilinearity. In addition to the shift-invariant trilinearity constraint a rank analysis step is used to access the number of principal components that is required to accurately reconstruct the amplitude spectra.

worth mentioning that SIT provides the second order advantage in situations where PARAFAC2 is troubled, specifically, if non-linear retention time shifts or even peak inversion occur, which are scenarios violating the cross-product constraint of PARAFAC2 [20,22]. A small example for second order calibration is given in the supporting information.

However, one limitation of the SIT model, is that the one-component PCA reconstruction of the amplitude spectra is only accurate as long as the distribution of frequencies contained in the time domain signal remains constant across the samples. Practically, that means that the SIT constraint is violated if the shape of the elution profiles changes across samples. Examples A and B in Fig. 3 show situations in which the SIT constraint holds, and C shows an example when SIT is violated. This is a matter of practical relevance when analytes show a concentration dependency in their elution behavior as described earlier. For these situations we propose a simple but effective extension of the SIT constraint that can accurately model changes in the elution profiles while maintaining the desirable properties of the SIT constraint.

The method we propose extends the SIT algorithm by a simple rank analysis step to determine the number of principal components that are required to reconstruct the amplitude spectra sufficiently accurate. The principle of the algorithm is exemplified in Fig. 4. The largest analyte peak of the yellow elution profiles has a broader peak width than the other analytes and subsequently a tighter amplitude spectrum. Hence, the matrix of amplitude spectra is a rank-two-system in which the first principal component is the average shape of the elution profiles, and the second principal component is a stretch or compression of the average shape. We call the new method shift-invariant soft trilinearity (SIST) as it allows for more flexibility than the shift-invariant trilinearity while being more rigid than non-negativity constraint MCR. On the other hand, SIST should perform similar to SIT if all elution profiles have the same shape and the data structure is shifted-trilinear. Applying SIST requires a definition of what a "sufficiently accurate reconstruction" means. Empirically, we found that a value of 99% explained variance works good for all our studied data sets, meaning that the number of components in the PCA model of the frequency domain data is chosen so that the variance explained is larger than 99%. Further, we observed that SIST is significantly faster if the rank analysis is performed after the algorithm has already converged to a certain point. Convergence is measured as the difference between the sum of squared residuals in iteration step *i* and iteration step *i*-1. The convergence criterion defines a numerical threshold to which the algorithm needs to converge to reach a solution. In our implementation the rank analysis is performed after the convergence is about 10 × convergence criterion.

3. Materials and methods

3.1. GC-MS data

The thermo-desorption (TDS) GC-MS data has been acquired in emission studies of 18 foam samples at BASF Polyurethanes GmbH Lemförde. The samples have been measured according to the VDA278-VOC guideline. The TDS-GC-MS measurements have been performed using a Shimadzu TD-20 thermo-desorption unit and a Shimadzu GC-2010 Plus gas chromatograph coupled to a Shimadzu GCMS-QP2020 mass spectrometer for detection. A TENAX TA trap was used to concentrate the VOC emissions in a micro chamber at 80 °C, 3% rel. Humidity and for a duration of 45 min. Two intervals shown in Fig. 5 were selected from the dataset to study the performance of SIST algorithm. The intervals are covering the retention time intervals 8.43–8.96 min and 36.97–37.21 min (see Fig. 5).

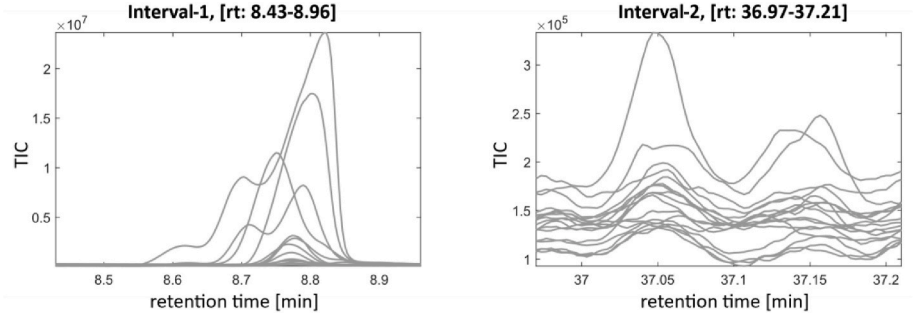


Fig. 5. Two selected intervals from the TDS-GC-MS data set which are studied to compare the performance of different chemometric methods.

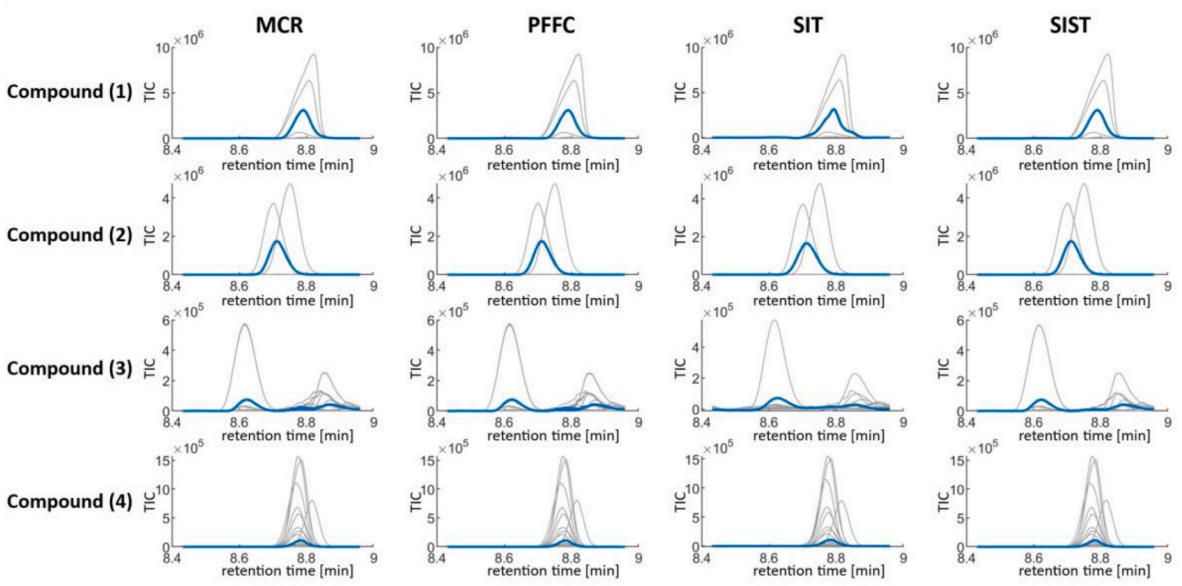


Fig. 6. Results of different decompositions used for modeling the second interval $I_{rt=8.43-8.96}$. **Grey colored:** Estimated elution profiles for all samples (results of the model that explained the original data best), **Blue colored:** Overlays of estimated elution profiles from ten repeated fits (top ten out of 50 repeated fits).

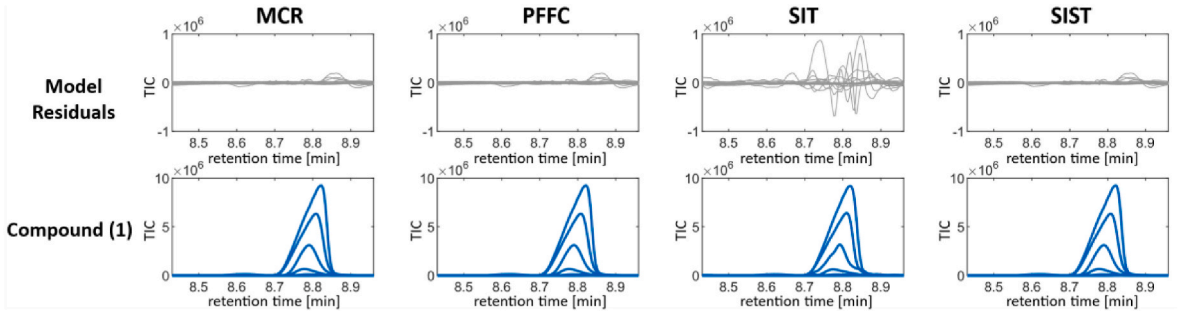


Fig. 7. Comparison of the model residuals and the factor resolution for analyte (1) eluting in the first retention time interval $I_{rt=8.43-8.96}$.

3.2. Data analysis

Data analysis was performed in MATLAB R2022b and in PARADISe v 6.0.0 [29]. The chromatographic data has been pre-processed using the alignment function implemented in PARADISe v 6.0.0., which is a combination of ico-shift and correlation optimized warping [30]. The characterization of the chemical compounds has been performed using the NIST11 database. The performance of SIST was evaluated in comparison to the performance of MCR, PARAFAC2 flexible coupling (PFFC), and SIT. The convergence criterion was set to 10^{-8} (fit difference between consecutive iterations) and the maximum number of iterations was set to 1000 for MCR, SIT and SIST and 5000 for PFFC. Only models that converged within the maximum number of iterations were considered. Further, non-negativity constraints were applied to all modes. For all methods, 50 repetitive fits to the respective intervals $I_{rt=8.43-8.96}$ and $I_{rt=36.97-37.21}$ have been executed, starting from different random initializations. For each method and interval, the ten models that reached the highest fit (variance in the original data explained by the model) are compared.

Beside a qualitative assessment of the resolved profiles, the evaluation specifically focused on 1) the variability of estimated factors across different models, and 2) the precision of factor estimates within the top ten models with the highest fit.

4. Results and discussion

The first interval $I_{rt=8.43-8.96}$ shows co-eluting peaks of (1) 1-(2-propenyloxy)-2-propanol, (2) chloro-benzene and (3) 4-methyl-2-hexanone. All chemometric methods (MCR, PFFC, SIT and SIST) could deconvolute the three co-eluting peaks and additionally separate the column signal of (4) hexamethyl-cyclotrisiloxane (Fig. 6).

The analysis of the model residuals, shown in Fig. 7, reveals that SIT fits the data worse than MCR, PFFC and SIST. The reason is that the elution profiles of (1) are showing a concentration dependent change of their shape. Hence, the SIT constraint is too rigid, and the estimates of the elution profiles are erroneous. Conversely, MCR, PFFC and SIST provide more accurate estimates of the true elution profiles which results in smaller residuals. The results show that SIST is flexible enough to model the shape changes that occur. Further, it can be emphasized that the precision of the factor estimates of the top ten out of 50 repeated fits is very high across all models, which is indicated by the blue lines in Fig. 6, showing the overlayed resolved elution profiles of the top ten models. Therefore, rotational ambiguity as well as local minima seem not to be a problem for this data set.

In the context of LC-MS, GC-MS, and LC-DAD, it has been reported in previous studies that PFFC and MCR have problems separating the baseline signal from very low intensity peaks [10,31,32]. Recently, Olivieri et al. investigated this problem and attributed it to the rotational ambiguity due to lacking selectivity in the elution profile mode [32]. In the same work, the authors proposed a background interpolation

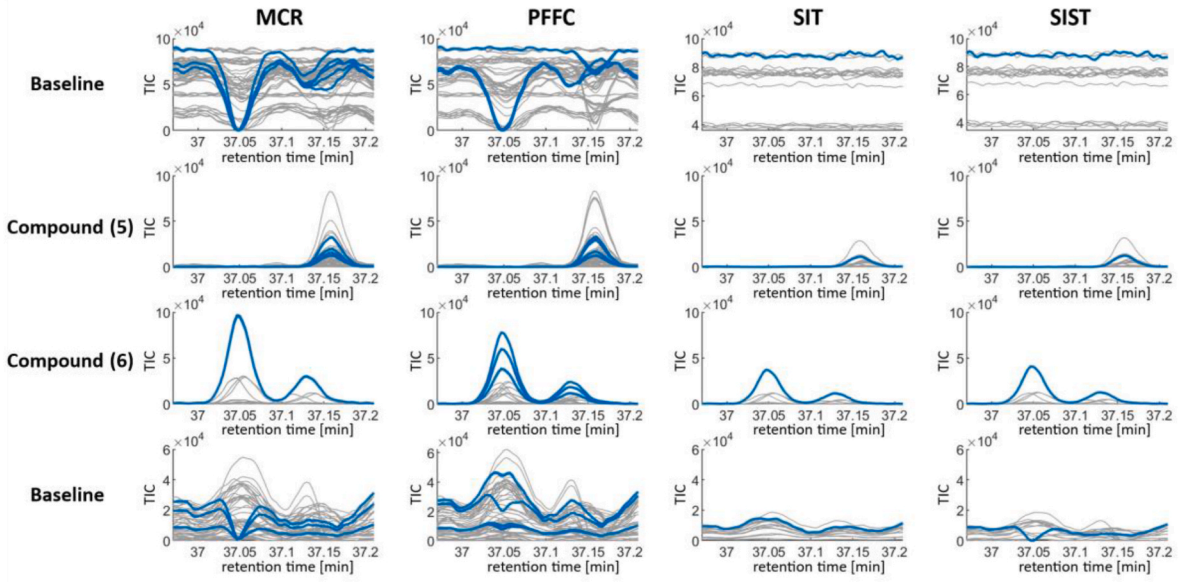


Fig. 8. Results of different decompositions used for modeling the second interval $I_{rt=36.97-37.21}$. **Grey colored:** Estimated elution profiles for all samples (results of the model that explained the original data best), **Blue colored:** Overlays of estimated elution profiles from ten repeated fits (top ten out of 50 repeated fits).

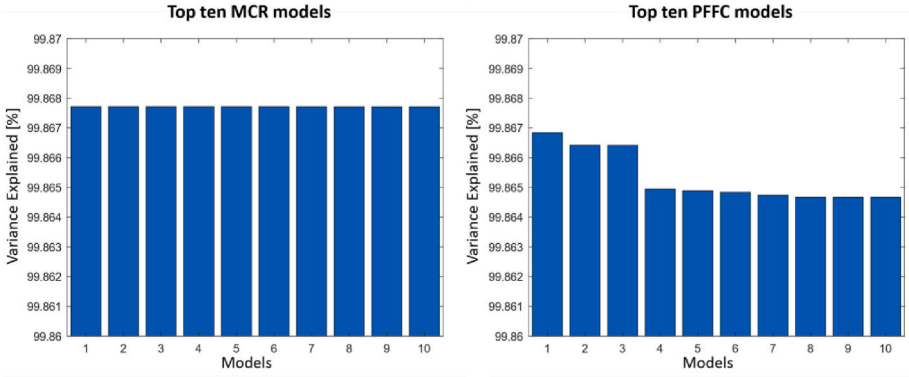


Fig. 9. Fit values for the top ten MCR models and the top ten PFFC models on the interval $I_{rt=36.97-37.21}$.

constraint for MCR as a remedy, which can effectively reduce the rotational ambiguity of the MCR solution. However, the background interpolation constraint requires *a priori* knowledge of the peak positions and peak widths of all analytes in a selected chromatographic interval [32]. In another study, Schneide et al. reported for GC-MS data that SIT can more effectively separate background from analyte signals than PFFC and PARARAC2, if the data follows a shifted-trilinear structure [8].

Hence, the second data set was selected in order to resemble a situation with a low intensity peak and a high baseline, to test if SIST can preserve the desirable properties of SIT despite its higher flexibility.

The second interval $I_{rt=36.97-37.21}$ shows co-eluting peaks of the two analytes (5) 2,4-bis(1,1-dimethylpropyl)-phenol and (6) tetramethyl-1,1′-biphenyl isomers (Fig. 8). To resolve the two analytes from baseline contributions, MCR, PFFC, SIT and SIST models were fitted. The SIST constraint was applied to the analyte and baseline components to test, whether the model would adjust the flexibility automatically. The concentration of the two analytes is very low compared to the high baseline signal. One can see that the results of SIT and SIST in Fig. 8 deviate from the results of MCR and PFFC. The reason is that MCR and PFFC have difficulties separating the baseline from the peak signals. The baseline signals estimated with SIT and SIST are flat compared to the curvier baseline estimates obtained with MCR and PFFC. This results in an overestimation of the peak areas of (5) and (6) for the elution profiles estimated with MCR and PFFC. Both, SIST and SIT provide overall very similar estimates for the elution profiles.

Another observation is the large variation in the factor estimates among the top ten out of 50 repeated fits for MCR and PFFC, in contrast to the high precision achieved with SIT and SIST. While it is known that non-negativity constrained MCR may suffer from rotational ambiguity without further constraints [16,17], no study has investigated the uniqueness properties of PFFC solutions. However, upon comparing the fit values of the top ten MCR models with the top ten PFFC models (Fig. 9), it appears plausible to assume different root causes for the observed variability in factor estimates for MCR and PFFC. The fit values of the top ten MCR models are essentially identical (based on the selected convergence criterion), suggesting that the variability can be attributed to rotational ambiguity. Conversely, the fit values of the top ten PFFC models show significant differences, indicating the presence of local minima. Since the SIST model is similar to the PARALIND model [33] in that the loading vectors in the time mode are in fact not vectors but span a low-dimensional subspace the results suggest that SIST has similar uniqueness properties. This would mean that the model is unique under mild conditions to the extent that the unfolded time loading stemming from the PCA model in Fig. 4 is unique (while the actual PCA components have rotational freedom). This was empirically verified by observing that even when the model is restarted multiple times from different random starting points, the solution is always the same (except for local minima). While this does not constitute a formal proof, it is

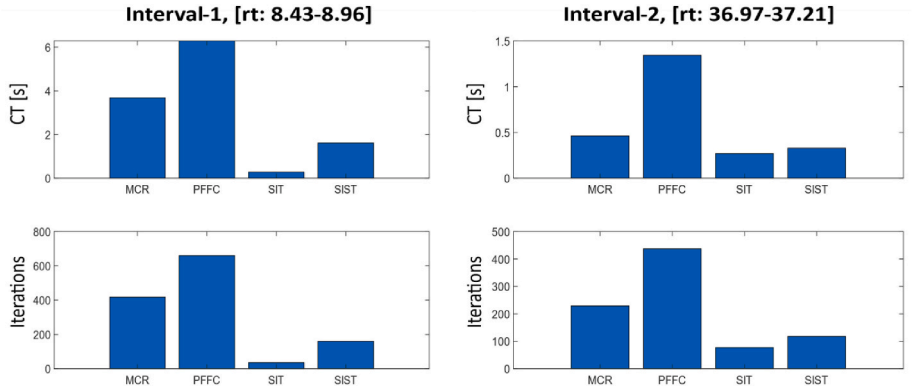


Fig. 10. Average computation time and average number of iterations until convergence for the top ten models for the two different intervals $I_{rt=8.43-8.96}$ and $I_{rt=36.97-37.21}$.

statistically unlikely that the repeated solutions end up the same if there is some rotational ambiguity.

The SIT and SIST algorithms demonstrate better convergence properties than PFFC as all solutions converged to the global optimum. In terms of computational costs, it is worth mentioning that the complexity added to the SIST algorithm, compared to SIT, makes SIST slightly more computationally heavy. Nevertheless, as shown in Fig. 10, SIT and SIST are both computationally cheaper than MCR and PFFC, mainly because they need fewer iterations to reach convergence, on the data sets investigated. In comparison, PFFC needs on average the most iterations and the longest computation time.

5. Conclusion

In conclusion, this study introduces the shift-invariant soft trilinearity (SIST) model as an effective solution for analyzing gas-chromatography coupled mass spectrometry data, particularly for addressing shifts and shape changes in elution profiles. Our findings demonstrate that SIST offers a balanced approach between the flexibility of bilinear models and the structure of trilinear constraints. Through empirical evaluation on real-world data, SIST seems to suffer substantially less from rotational ambiguity than non-negativity constrained MCR and suffers less from convergence problems than PARAFAC2 flexible coupling (PFFC). We hypothesize that the factors resolved by the SIST model may be uniquely identified by the subspace that the principal components span. However, a thorough study investigating the area of feasible solutions [34,35] is required to test this hypothesis. Further, the proposed method extends the flexibility of the SIT model to deal with shape changes in elution profiles across samples. The key innovation of SIST lies in its dynamic adjustment between bilinear and trilinear modeling, providing a tailored approach to chromatographic data analysis aligning closely with the actual structure of the data. However, it is important to note that the proposed method introduces a new model parameter that needs eventual tuning. While for the data sets investigated, a variance explained criterion worked fine for adjusting the model flexibility, it is not guaranteed that this criterion can be used universally. Data sets with a different noise structure (e.g. GC-TOF-MS or LC-HRMS) may require a new tuning of the flexibility parameter. Hence, future work may focus on further validating the model across a broader range of data sets and exploring its applicability to other analytical techniques.

CRediT authorship contribution statement

Paul-Albert Schneide: Writing – original draft, Funding acquisition, Formal analysis, Conceptualization. **Neal B. Gallagher:** Writing – original draft, Supervision, Conceptualization. **Rasmus Bro:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

The authors would like to acknowledge the support of Marc Bockhoff and Philipp Rotering from BASF Polyurethanes GmbH Lemförde in providing the TDS-GC-MS data and the method description. Furthermore, the authors would like to thank the two anonymous reviewers for constructive feedback and fruitful discussions. We would like to thank one anonymous reviewer for providing a simulated data set giving us the opportunity to objectively test the performance of our methods.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105155.

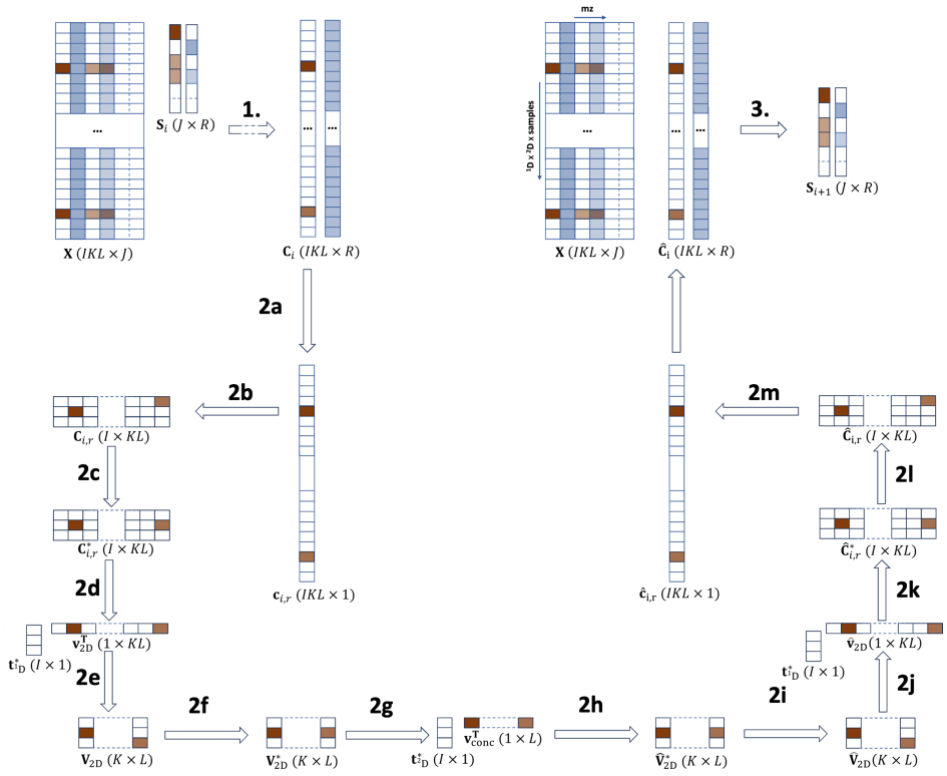
References

[1] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309.
[2] T. Skov, R. Bro, Solving fundamental problems in chromatographic analysis, Anal. Bioanal. Chem. 390 (2008) 281–285.
[3] X. Fan, et al., Deep-Learning-Assisted multivariate curve resolution, J. Chromatogr. A 1635 (2021).
[4] H. Parastar, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chem. 86 (2014) 286–297.
[5] Y.Y. Chang, et al., Comparison of three chemometric methods for processing HPLC-DAD data with time shifts: simultaneous determination of ten molecular targeted anti-tumor drugs in different biological samples, Talanta 224 (2021).
[6] C. Pérez-López, B. Oró-Nolla, S. Lacorte, R. Tauler, Regions of interest multivariate curve resolution liquid chromatography with data-independent acquisition tandem mass spectrometry, Anal. Chem. 95 (2023) 7519–7527.
[7] G. Baccolo, B. Quintanilla-Casas, S. Vichi, D. Augustijn, R. Bro, From untargeted chemical profiling to peak tables – a fully automated AI driven approach to untargeted GC-MS, TrAC, Trends Anal. Chem. 145 (2021) 116451.
[8] P.A. Schneide, R. Bro, N.B. Gallagher, Shift-invariant tri-linearity—a new model for resolving untargeted gas chromatography coupled mass spectrometry data, J. Chemom. 37 (2023).
[9] R. Tauler, Multivariate curve resolution of multiway data using the multilinearity constraint, J. Chemom. 35 (2021).
[10] X. Zhang, R. Tauler, Flexible implementation of the trilinearity constraint in multivariate curve resolution alternating least squares (MCR-ALS) of chromatographic and other type of data, Molecules 27 (2022) 2338.
[11] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemom. 13 (1999) 275–294.

[12] Jeremy E. Cohen, R. Bro, Nonnegative PARAFAC2: a flexible coupling approach, in: Yannick Deville, S. Gannot, M. R, P. M. D, W. D (Eds.), Latent Variable Analysis and Signal Separation, Springer International Publishing, Cham, 2018, pp. 89–98.
[13] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – a review, Anal. Chim. Acta 1145 (2021) 59–78.
[14] O.S. Borgen, B.R. Kowalski, An extension of the multivariate component-resolution method to three components, Anal. Chim. Acta 174 (1985) 1–26.
[15] W.H. Lawton, E.A. Sylvestre, Self modeling curve resolution, Technometrics 13 (1971) 617–633.
[16] M. Sawall, H. Schröder, D. Meinhardt, K. Neymeyr, On the ambiguity underlying multivariate curve resolution methods, in: Comprehensive Chemometrics, Elsevier, 2020, pp. 199–231, https://doi.org/10.1016/B978-0-12-409547-2.14582-2.
[17] A.C. Olivieri, A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution – a tutorial, Anal. Chim. Acta 1156 (2021) 338206.
[18] J.R. Conder, Peak distortion in chromatography. Part 1: concentration-dependent behavior, J. High Resolut. Chromatogr. 5 (1982) 341–348, https://doi.org/10.1002/jhrc.1240050702. Preprint at.
[19] Harshman, R. A. NOTE: This Manuscript Was Originally Published in 1972 and Is Reproduced Here to Make it More Accessible to Interested Scholars. The Original Reference Is Harshman, RA (1972). PARAFAC2: Mathematical and Technical Notes. UCLA Working Papers in Phonetics, 22, 30-44.(University Microfilms, Ann Arbor, Michigan, No. 10,085).
[20] S.A. Bortolato, A.C. Olivieri, Chemometric processing of second-order liquid chromatographic data with UV-vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, Anal. Chim. Acta 842 (2014) 11–19.
[21] J.M.F. Ten Berge, H.A.L. Kiers, Some uniqueness results for PARAFAC2, Psychometrika 61 (1996).
[22] M.B. Anzardi, J.A. Arancibia, A.C. Olivieri, Interpretation of matrix chromatographic-spectral data modeling with parallel factor analysis 2 and multivariate curve resolution, J. Chromatogr. A 1604 (2019).
[23] T. Diera, et al., A non-target screening study of high-density polyethylene pipes revealed rubber compounds as main contaminant in a drinking water distribution system, Water Res. 229 (2023) 119480.
[24] J.M. Amigo, et al., Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis, J. Chromatogr. A 1217 (2010) 4422–4429.
[25] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with PARAFAC2, TrAC, Trends Anal. Chem. 27 (2008) 714–725.
[26] H.E. Toraman, et al., Application of Py-GC/MS coupled with PARAFAC2 and PLS-DA to study fast pyrolysis of genetically engineered poplars, J. Anal. Appl. Pyrolysis 129 (2018) 101–111.
[27] H. Yu, R. Bro, N.B. Gallagher, PARASIAS: a new method for analyzing higher-order tensors with shifting profiles, Anal. Chim. Acta 1238 (2023) 339848.
[28] Fourier and wavelet transforms, in: Data-Driven Science and Engineering 47–83, Cambridge University Press, 2019, https://doi.org/10.1017/9781108380690.003.
[29] R.B.J.L.H. Beatriz Quintanilla-Casas, et al., Tutorial on PARADISe: PARAFAC2-Based Deconvolution and Identification System for Processing GC–MS Data, 2023.
[30] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (2004) 231–241.
[31] O.M. Kronik, X. Liang, N.J. Nielsen, J.H. Christensen, G. Tomasi, Obtaining clean and informative mass spectra from complex chromatographic and high-resolution all-ions-fragmentation data by nonnegative parallel factor analysis 2, J. Chromatogr. A 1682 (2022) 463501.
[32] M.D. Carabajal, R.P. Vidal, J.A. Arancibia, A.C. Olivieri, A new constraint to model background signals when processing chromatographic-spectral second-order data with multivariate curve resolution, Anal. Chim. Acta 1266 (2023).
[33] R. Bro, R.A. Harshman, N.D. Sidiropoulos, M.E. Lundy, Modeling multi-way data with linearly dependent loadings, J. Chemom. 23 (2009) 324–340.
[34] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemometr. Intell. Lab. Syst. 103 (2010) 96–107.
[35] M. Sawall, K. Neymeyr, A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: theoretical foundation, inverse polygon inflation, and FAC-PACK implementation, J. Chemom. 28 (2014) 633–644.

Paper 3

Schneide P-A, Armstrong MS, Gallagher NB, Bro R. Unlocking new capabilities in the analysis of GC×GC-TOFMS data with shift-invariant multi-linearity. Journal of Chemometrics (submitted)



Journal of CHEMOMETRICS

Unlocking new capabilities in the analysis of GC×GC-TOFMS data with shift-invariant multi-linearity

Journal:	*Journal of Chemometrics*
Manuscript ID	CEM-24-0086.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	07-Jul-2024
Complete List of Authors:	Schneide, Paul-Albert; Kobenhavns Universitet Sektion for Bioinformatik og RNA Biologi; BASF SE Armstrong, Michael Sorochan; Universidad de Granada Departamento de Linguistica General y Teoria de la Literatura Gallagher, Neal; Eigenvector Research Inc Bro, Rasmus; Kobenhavns Universitet Sektion for Bioinformatik og RNA Biologi
Keyword:	Shift-invariant tensor decomposition, GC×GC-TOFMS, MCR

Unlocking new capabilities in the analysis of GC×GC-TOFMS data with shift-invariant multi-linearity

Authors

Paul-Albert Schneide[1,2], Michael Sorochan Armstrong[3], Neal Gallagher[4], Rasmus Bro[1]

[1] Department of Food Science, University of Copenhagen, Frederiksberg, Denmark

[2] Department of Analytical Science, BASF SE, Ludwigshafen am Rhein, Rhineland-Palatinate, Germany

[3] Departamento de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, Spain

[4] Eigenvector Research, Inc., Manson, Washington, USA

Abstract

This paper introduces a novel deconvolution algorithm, shift-invariant multi-linearity (SIML), which significantly enhances the analysis of data from two-dimensional gas chromatography instruments coupled to a time-of-flight mass spectrometer (GC×GC-TOFMS). Designed to address the challenges posed by retention time shifts and high noise levels, SIML incorporates wavelet-based smoothing and Fourier-Transform based shift-correction within the multivariate curve resolution-alternating least squares (MCR-ALS) framework. We benchmarked the SIML algorithm against non-negativity constrained MCR-ALS and Parallel Factor Analysis 2 with flexible coupling (PARAFAC2×N) using both simulated and real GC×GC-TOFMS datasets. Our results demonstrate that SIML provides unique solutions with significantly improved robustness, particularly in low signal-to-noise ratio scenarios, where it maintains high accuracy in estimating mass spectra and concentrations. The enhanced reliability of quantitative analyses afforded by SIML underscores its potential for broad application in complex matrix analyses across environmental science, food science, and biological research.

Keywords

Shift-invariant tensor decomposition, GC×GC-TOFMS, MCR

Abbreviations

^{1}D	First retention dimension in two-dimensional chromatography
^{2}D	Second retention dimension in two-dimensional chromatography
ALS	Alternating Least Squares
BPC	Base peak chromatogram
CLS	Classical least squares
FFT	Fast Fourier Transform
FT	Fourier Transform
GC × GC-TOFMS	Two dimensional gas chromatography coupled to a mass spectrometric (time-of-flight) detector
MCR	Multivariate curve resolution
NMF	Non-negative matrix factorization
PARAFAC	Parallel factor analysis
PARAFAC2	Parallel factor analysis 2
PARAFAC2xN	Parallel factor analysis 2 with flexible coupling constraint allowing for shift in more than one mode
SIML	Shift-invariant multi-linearity algorithm
SIML-DN	Shift-invariant multi-linearity algorithm with denoising
SNR	Signal-to-noise ratio
SSE	Sum of squared errors
SST	Total sum of squares
TIC	Total ion chromatogram
SVD	Singular value decomposition
TMS	Tri-methyl-silyl (protective group)

Notation

$\underline{\mathbf{X}}^{(p)}$	*pth* order tensor
$\mathbf{X}$	$(I \times J)$ matrix, equivalent to $\underline{\mathbf{X}}^{(2)}$
$\mathbf{x}$	$(I \times 1)$ vector, equivalent to $\underline{\mathbf{X}}^{(1)}$
x_{ijk}	Scalar, equivalent to $\underline{\mathbf{X}}^{(0)}$
$\mathbf{x}^{\mathrm{T}}$	$(I \times 1)$ transpose of $\mathbf{x}$ $(1 \times I)$
$\hat{\mathbf{X}}$	estimate of $\mathbf{X}$
$\mathbf{C}^{*}$	amplitude spectra of $\mathbf{C}$ after 1D-FFT has been applied column-wise
$\|\| \cdot \|\|_F^2$	Frobenius Norm

1. Introduction

Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC-TOFMS) is a powerful analytical technique that allows for comprehensive separation and characterization of very complex samples such as pyrolysis oils, environmental samples, biological, and food samples. [1–5] However, handling the data is challenging and time consuming. The challenge is particularly evident in exploratory, untargeted analysis where the aim is to obtain an exhaustive chemical fingerprint of the sample composition.[6–8] A recently published benchmark of eight different commercial and open-source software packages revealed larger differences in the data processing capabilities of the individual software packages, most notably with respect to the identification of features in untargeted analysis.[9] In addition to the missing standardization in the data processing workflows of available software tools, their functionality is in some cases not sufficient, which motivates the development of novel algorithms for more sophisticated chemical information extraction (deconvolution), pattern recognition, or downstream statistical analysis.[7,8,10–13] Chemometric methods such as Multivariate Curve Resolution [MCR[14–16], if non-negativity constrained it can also be referred to as non-negative matrix factorization (NMF)[17–19] in the data science community], Parallel Factor Analysis (PARAFAC) and extended versions of Parallel Factor Analysis 2 (PARAFAC2×N) have been described as useful methods for deconvolution in targeted and untargeted GC × GC-TOFMS data analysis. [13,20,21] Nevertheless, there are limitations associated with each of the currently known deconvolution algorithms in their application to GC × GC-TOFMS data analysis. For example, the structure of the MCR model accounts for retention time shifts occurring in the first and second retention dimension; however, MCR does not generally provide unique solutions.[22,23] The rotational ambiguity of MCR solutions can cause large variabilities in the estimated qualitative and quantitative information.[24] On the other hand, PARAFAC and PARAFAC2 provide unique solutions but have higher requirements in regard to the data structure. In the case of PARAFAC, the GC × GC-TOFMS data would need to be perfectly aligned to maintain the assumption of multi-linearity. Unfortunately, retention time shifts violate the PARAFAC model assumptions.[25] The PARAFAC2×N method is more flexible and can model data that deviates from a multi-linear structure caused by retention time shifting but comes with a high algorithmic complexity.[13,26]

In this paper a new deconvolution algorithm called shift-invariant multi-linearity (SIML) is presented. The algorithm is inspired by the multi-linearity constraint proposed by Tauler et al. and the shift-invariant tri-linearity constraint proposed by Schneide et al.[27,28] The proposed method integrates a wavelet-based smoothing and the shift-invariance properties of the Fourier-Transform into the MCR-ALS routine to effectively correct intra- and inter-sample shifts yielding unique solutions. The algorithm is benchmarked

against MCR-ALS with non-negativity constraints and PARAFAC2×N on challenging simulated and real multi-sample GC × GC-TOFMS data sets.

2. Background

2.1 Data structure

The data structure of a single GC-MS measurement can be described as a matrix $\mathbf{X}$ with dimensions $(I \times J)$, with I denoting the number of scans in the first retention dimension (^{1}D) and J denoting the mass scans (mz). In extension to that, a GC × GC-TOFMS measurement naturally has the form of a third order tensor $\underline{\mathbf{X}}^{(3)}$ with dimensions $(I \times K \times J)$ in which K describes the scans in the second retention dimension (^{2}D). One practical way of visualizing $\underline{\mathbf{X}}^{(3)}$ is thinking of it as K slices $\underline{\mathbf{X}}^{(2)}$ with dimensions $(I \times J)$ or as I slices $\underline{\mathbf{X}}^{(2)}$ of dimension $(K \times J)$. Thus, a GC × GC-TOFMS measurement can also be expressed in the form of K concatenated slices $\underline{\mathbf{X}}^{(2)}$ with dimensions $(I \times J)$, giving an augmented matrix $\mathbf{X}$ with dimensions $(IK \times J)$. This idea also extends to the situation of having a set of several GC × GC-TOFMS measurements, which can be arranged to form a fourth order tensor $\underline{\mathbf{X}}^{(4)}$ with dimensions $(I \times K \times L \times J)$ or an augmented matrix $\mathbf{X}$ with dimensions $(IKL \times J)$, where L is the number of samples. To aid visualization, examples of the data structures for the case of a single GC × GC-TOFMS measurement and multiple GC × GC-TOFMS measurements are given in Figure 1. An example peak location is indicated by the red boxes and is shown to shift for multiple samples (Figure 1B). The blue boxes in Figure 1 are indicating a background signal.

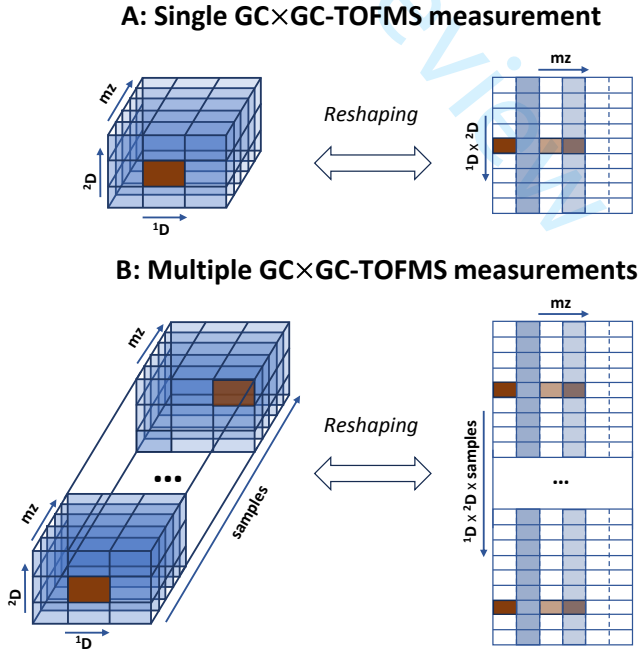


*Figure 1: Visualization of the data structure of **A:** a single GC × GC-TOFMS measurement organized as higher order tensor or as augmented matrix and **B:** a set of multiple GC × GC-TOFMS measurements organized as higher order tensor or as augmented matrix.*

2.2 Algorithms for modeling GC×GC-TOFMS data

Different chemometric approaches have been described for extracting quantitative (concentrations) and qualitative (mass spectra) information from GC × GC-TOFMS data.[7,11–13,20,21] Specifically, MCR-ALS, PARAFAC and PARAFAC2×N were described in literature and will be explained in the following section.[13,20,21]

Multivariate Curve Resolution is a bilinear factorization method which decomposes a matrix $\mathbf{X}$ into two positive matrices $\mathbf{C}$ and $\mathbf{S}$. In the context of GC × GC-TOFMS data analysis, $\mathbf{X}$ is the set of unfolded GC × GC-TOFMS measurements with dimensions $(IKL \times J)$, $\mathbf{C}$ is a factor matrix with dimensions $(IKL \times R)$ containing the concatenated elution profiles, and $\mathbf{S}$ is a $(J \times R)$ sized factor matrix holding the analyte mass spectra. The MCR model can be formulated according to Equation 1:

Equation 1:
$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E}$$

Different algorithms have been proposed to calculate $\mathbf{C}$ and $\mathbf{S}$ but this paper will focus on the most prominently used Alternating Least Squares (ALS) algorithm. The loss function for the ALS algorithm with non-negativity constraints[29] can be formulated according to Equation 2:

Equation 2:
$$L(\mathbf{C},\mathbf{S}) = ||\mathbf{X} - \mathbf{C}\mathbf{S}^{\mathrm{T}}||_{\boldsymbol{F}}^{\boldsymbol{2}} \; s.t.$$
$$\mathbf{C}_{nr} \geq 0 \; \forall \; n \in \{1,\ldots, IKL\}, r \in \{1,\ldots, R\}$$
$$\mathbf{S}_{mr} \geq 0 \; \forall \; m \in \{1,\ldots, J\}, r \in \{1,\ldots, R\}$$

A major advantage of MCR is that it can deconvolve overlapped signals and effectively model chromatographic artifacts such as retention time shifts and changes in peak shape. But a major disadvantage of MCR is that it suffers from a rotational ambiguity which means that a range of solutions for $\mathbf{C}$ and $\mathbf{S}$ exist that all minimize $L(\mathbf{C},\mathbf{S})$. Mathematically, this can be shown by Equation 3 in which $\mathbf{C}$ and $\mathbf{S}$ are solutions obtained by fitting the MCR model with one set of initial values and $\mathbf{C}_{\mathrm{A}}$ and $\mathbf{S}_{\mathrm{A}}$ are rotated solutions (fulfilling the applied constraints) that would provide the same fit. The matrix $\mathbf{T}$ is a rotation matrix, such that the matrix product $\mathbf{T}\mathbf{T}^{-\mathbf{1}}$ becomes the identity matrix with dimensions equal to the number of modeled components.

Equation 3:
$$\mathbf{C}_{\mathrm{A}}\mathbf{S}_{\mathrm{A}}^{\mathrm{T}} = (\mathbf{C}\mathbf{T})(\mathbf{T}^{-\mathbf{1}}\mathbf{S}^{\mathrm{T}})$$

Several scientific works have investigated methods for estimating the range of feasible solutions theoretically or practically to derive estimates on how well-defined a given solution is.[22,24,30,31] Rotational ambiguity can be reduced by constraining $L(\mathbf{C},\mathbf{S})$ and incorporating a priori knowledge about the measurement principle and resulting characteristics of the data (unimodal shape, selectivity, local rank or multi-linearity).[32–34] Specifically, the multi-linearity constraint[27] can provide unique solutions for data structures $\underline{\mathbf{X}}^{(p\geq 3)}$ (e.g., GC × GC-TOFMS data)[27], and can be seen as a particular implementation of the PARAFAC/CANDECOMP model.[35,36] However, multi-linearity is a strong constraint that requires elution profiles of a given analyte to remain constant in shape and position across scans within a sample, and sample-to-sample.[25] These conditions for multi-linearity are not always satisfied for GC × GC-TOFMS data, most importantly because of retention time shifts. Although the application of PARAFAC for the decomposition of single GC × GC-TOFMS measurements has been described in literature[21], this approach will only be valid under specific experimental conditions limiting the operation range of chromatographic methods.[7] For instance, if the temperature in ^{2}D is not held constant alongside increasing scan number in ^{1}D, retention times in ^{2}D will be shorter because analytes are reaching the detector earlier.[20] This case is referred to as intra sample shift and the effect is shown in Figure 3A. [20] Consequently, the PARAFAC model will be biased and more flexible alternatives such as flexible-coupling-PARAFAC2[13,37] or shift-invariant tri-linearity constrained MCR will likely give more accurate results.[28,38]

The situation becomes more complicated if the goal is to analyze a set of several GC × GC-TOFMS measurements jointly. In this situation, additional random shifts in ^{1}D and ^{2}D occur together with the intra-sample shifts in ^{2}D (Figure 3B). Thus, shifts in both retention dimensions need to be accounted for.[20] To handle deviations from a multi-linear data structure, the PARAFAC2×N algorithm has been published as an extension to the flexible-coupling-PARAFAC2 model, which can effectively handle shifts in both retention dimensions.[13] However, the way algorithms from the "PARAFAC2-family" fundamentally address deviations from multi-linearity is rather complicated and come with assumptions regarding the nature of the shifts occurring in ^{1}D and ^{2}D that may have practical implications (see Supporting Information 1).[39,40] Therefore, a novel shift-invariant multilinearity constraint is proposed to account for shifts in both retention dimensions while providing unique solutions.

Conceptually, the shift-invariant multi-linearity constraint utilizes the shift-invariant property of the Fourier Transform modulus (amplitude spectra) to "de-shift" the estimates of elution profiles within the MCR-ALS routine to transform a non-multi-linear problem into a multi-linear problem, for which a unique solution exists.[28] Another approach for synchronizing factor estimates inside the MCR-ALS routine based on a peak alignment strategy has been proposed by Zhang et al.[38]

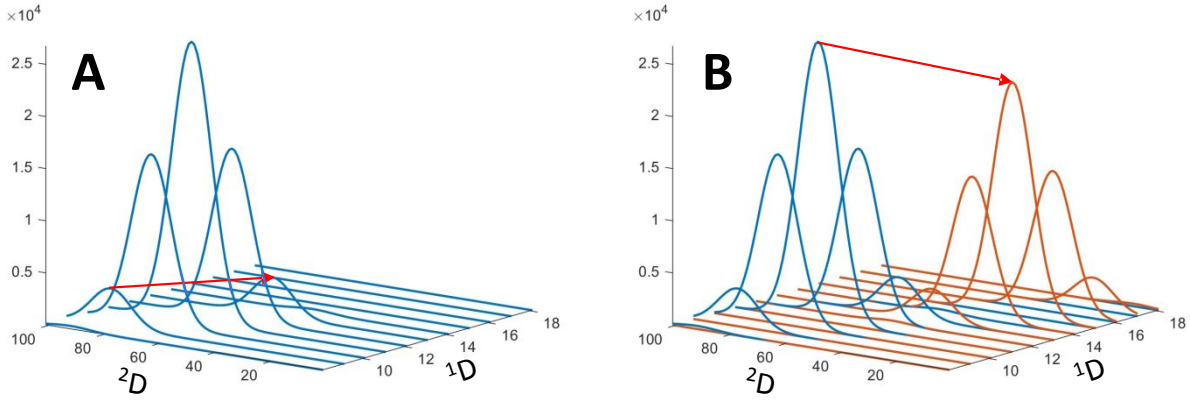


Figure 2: Visualized are TIC profiles of one analyte present in two different samples (blue and orange peaks). ***A:*** *Retention in D^2 gets reduced with increasing scan number in D^1 because of higher temperatures on the second column (intra sample shift).* ***B:*** *Retention times between samples (plotted together as overlay) can vary because chromatographic conditions cannot be kept perfectly constant (inter sample shift).*

2.3 Shift-invariant multi-linearity

The shift-invariant multi-linearity constraint is implemented in the MCR-ALS routine as schematically shown in Figure 3. The example shows a simple case of a two-component system consisting of an analyte being present in multiple GC × GC-TOFMS measurements and a baseline signal. The steps for applying shift-invariant multi-linearity follow:

Step 1: Regressing spectral estimates $\mathbf{S}_i$ $(J \times R)$ onto $\mathbf{X}$ $(IKL \times J)$ in a non-negative classical least squares (CLS) step to obtain estimates of concatenated elution profiles $\mathbf{C}_i$ $(IKL \times R)$. [41] With R denoting the number of components and i denoting the iteration count. If $i = 0$, $\mathbf{S}_0$ contains initial values, e.g., random positive numbers or more sophisticated initial estimates. [42,43] If $i > 0$, $\mathbf{S}_i$ is the output from step 3 of the previous iteration $i - 1$.

Input: Data matrix $\mathbf{X}$ $(IKL \times J)$ and spectral estimates $\mathbf{S}_i$ $(J \times R)$,

Output: $\mathbf{C}_i$ $(IKL \times R)$

Step 2: Applying the shift-invariant multi-linearity constraint and optionally smoothing to the concatenated elution profiles by iterating over all $r \in \{1,\ldots, R\}$ columns of $\mathbf{C}_i$, and applying the procedures described in sub steps 3a-m, to obtain the constrained, concatenated elution profiles $\hat{\mathbf{C}}_\mathrm{i}$ $(IKL \times R)$

Input: $\mathbf{C}_i$ $(IKL \times R)$

Output: $\hat{\mathbf{C}}_\mathrm{i}$ $(IKL \times R)$

Sub step 2 a (optional): Wavelet based denoising as described in the supporting information (SI 2.1 and 2.2).

Input: $\mathbf{c}_{i,r}$ $(IKL \times 1)$

Output: $\mathbf{c}_{i,r}$ $(IKL \times 1)$

Sub step 2 b: Reshaping $\mathbf{c}_{i,r}$ $(IKL \times 1)$ to $\mathbf{C}_{i,r}$ $(I \times KL)$

Input: $\mathbf{c}_{i,r}$ $(IKL \times 1)$

Output: $\mathbf{C}_{i,r}$ $(I \times KL)$

Sub step 2 c: FFT is performed on the columns of $\mathbf{C}_{i,r}$ to synchronize elution profile estimates of the first retention dimension.

Input: $\mathbf{C}_{i,r}$ $(I \times KL)$

Output: $\mathbf{C}^*_{i,r}$ $(I \times KL)$

Sub step 2 d: One-component-SVD filter is applied to $\mathbf{C}^*_{i,r}$ to enforce multi-linearity to the amplitude spectra of the first retention dimension elution profiles $\mathbf{t}^*_{1\mathrm{D}}$ $(I \times 1)$. The concatenated elution profiles of the second retention dimension are obtained in $\mathbf{v}^{\mathbf{T}}_{2\mathrm{D}}$ $(1 \times KL)$.

Input: $\mathbf{C}^*_{i,r}$ $(I \times KL)$

Output: $\mathbf{t}^*_{1\mathrm{D}}$ $(I \times 1)$, $\mathbf{v}^{\mathbf{T}}_{2\mathrm{D}}$ $(1 \times KL)$

Sub step 2 e: Reshaping $\mathbf{v}^{\mathbf{T}}_{2\mathrm{D}}$ $(1 \times KL)$ to $\mathbf{V}_{2\mathrm{D}}$ $(K \times L)$

Input: $\mathbf{v}^{\mathbf{T}}_{2\mathrm{D}}$ $(1 \times KL)$

Output: $\mathbf{V}_{2\mathrm{D}}$ $(K \times L)$

Sub step 2 f: FFT is performed on the columns of $\mathbf{V}_{2\mathrm{D}}$ to synchronize elution profile estimates of the second retention dimension.

Input: $\mathbf{V}_{2\mathrm{D}}$ $(K \times L)$

Output: $\mathbf{V}^*_{2\mathrm{D}}$ $(K \times L)$

Sub step 2 g: One-component-SVD filter is applied to enforce multi-linearity to the amplitude spectra of the second retention dimension elution profiles $\mathbf{t}_{2\mathrm{D}}^{*}$ $(I \times 1)$. The relative concentrations of the modeled analyte in each of the concatenated samples is obtained in $\mathbf{v}_{\mathrm{conc}}^{\mathbf{T}}$ $(1 \times L)$

Input: $\mathbf{V}_{2\mathrm{D}}^{*}$ $(K \times L)$

Output: $\mathbf{t}_{2\mathrm{D}}^{*}$ $(I \times 1)$, $\mathbf{v}_{\mathrm{conc}}^{\mathbf{T}}$ $(1 \times L)$

Sub step 2 h: Estimating $\hat{\mathbf{V}}_{2\mathrm{D}}^{*}$ $(K \times L)$ from the results of the one-component SVD.

Input: $\mathbf{t}_{2\mathrm{D}}^{*}$ $(I \times 1)$, $\mathbf{v}_{\mathrm{conc}}^{\mathbf{T}}$ $(1 \times L)$

Output: $\hat{\mathbf{V}}_{2\mathrm{D}}^{*}$ $(K \times L)$

Sub step 2 i: inverse-FFT is performed on the columns of $\hat{\mathbf{V}}_{2\mathrm{D}}^{*}$ $(K \times L)$ to restore the shift-invariant-multi-linearity constrained estimates of the elution profiles (^{2}D) in the time domain.

Input: $\hat{\mathbf{V}}_{2\mathrm{D}}^{*}$ $(K \times L)$

Output: $\hat{\mathbf{V}}_{2\mathrm{D}}(K \times L)$

Sub step 2 j: Reshaping $\hat{\mathbf{V}}_{2\mathrm{D}}(K \times L)$ to $\hat{\mathbf{v}}_{2\mathrm{D}}(1 \times KL)$

Input: $\hat{\mathbf{V}}_{2\mathrm{D}}(K \times L)$

Output: $\hat{\mathbf{v}}_{2\mathrm{D}}(1 \times KL)$

Sub step 2 k: Estimating $\hat{\mathbf{C}}_{i,r}^{*}$ $(I \times KL)$ from the results of the one-component SVD.

Input: $\mathbf{t}_{1\mathrm{D}}^{*}$ $(I \times 1)$, $\hat{\mathbf{v}}_{2\mathrm{D}}(1 \times KL)$

Output: $\hat{\mathbf{C}}_{i,r}^{*}$ $(I \times KL)$

Sub step 2 l: inverse-FFT is performed on the columns of $\hat{\mathbf{C}}_{i,r}^{*}$ $(I \times KL)$ to restore the shift-invariant-multi-linearity constrained estimates of the elution profiles (^{1}D) in the time domain

Input: $\hat{\mathbf{C}}_{i,r}^{*}$ $(I \times KL)$

Output: $\hat{\mathbf{C}}_{\mathrm{i,r}}$ $(I \times KL)$

Sub step 2 m: reshaping $\hat{\mathbf{C}}_{\mathrm{i,r}}$ $(I \times KL)$ to $\hat{\mathbf{c}}_{\mathrm{i,r}}$ $(IKL \times 1)$

Input: $\hat{\mathbf{C}}_{\mathrm{i,r}}$ $(I \times KL)$

Output: $\hat{\mathbf{c}}_{\mathrm{i,r}}$ $(IKL \times 1)$

Step 3: Regressing $\hat{\mathbf{C}}_i$ $(IKL \times R)$ onto $\mathbf{X}$ in a non-negative classical least squares (CLS) step to obtain estimates of concatenated elution profiles $\mathbf{S}_{i+1}$ $(J \times R)$. The procedure terminates if changes in $L(\mathbf{C},\mathbf{S})$ are smaller than the defined convergence criterion or if i equals the defined maximum number of iterations.

Input: $\mathbf{C}_i$ $(IKL \times R)$

Output: $\hat{\mathbf{C}}_{\mathrm{i}}$ $(IKL \times R)$

Although, the denoising step in 2a is not strictly required for the multi-linearity, results discussed below show that it ensures very accurate estimates of mass spectra and elution profiles even at very low signal-to-noise ratios (SNR). Applying smoothing to the estimated elution profiles inside the ALS routine was found to be advantageous compared to smoothing of the raw data (compare Supporting Information 2). The combination of shift-invariant multi-linearity and wavelet based denoising will be distinguished from shift-invariant multi-linearity (SIML) by the abbreviation SIML-DN.

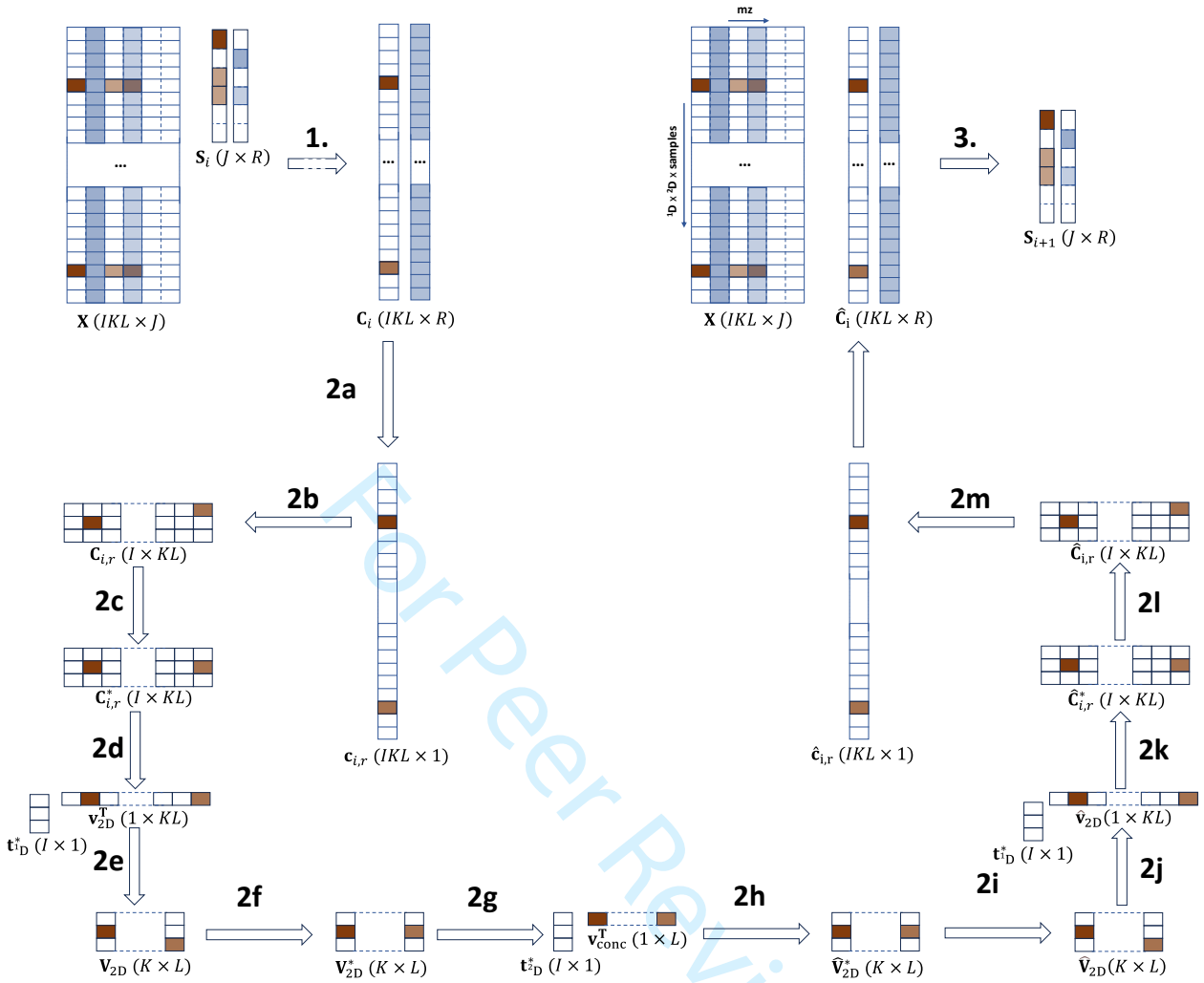


Figure 3: Visualization of one iteration cycle of the shift-invariant multi-linearity algorithm with denoising. In step 1, the concatenated elution profiles $\boldsymbol{C}_i\ (IKL \times R)$ are estimated from the data matrix $\boldsymbol{X}\ (IKL \times J)$ and the spectral estimates $\boldsymbol{S}_i\ (J \times R)$. In step 2, $\boldsymbol{C}_i\ (IKL \times R)$ is constrained to be shift-invariant multi-linear and the output from this step $\hat{\boldsymbol{C}}_i\ (IKL \times R)$. In step 3, $\boldsymbol{S}_{i+1}$ $(J \times R)$ is estimated from $\hat{\boldsymbol{C}}_i\ (IKL \times R)$ and $\boldsymbol{X}$.

3. Material and methods

3.1 Software and algorithms

The performance of shift-invariant multi-linearity with and without denoising was benchmarked against the performance of non-negativity constrained MCR-ALS, and PARAFAC2×N. The same criterion for algorithmic convergence was applied, based on the relative change in the loss function value. The maximum number of iterations was set to 1.000 and it was assured that only converged models were used in the benchmark. All algorithms were initialized with positive random values, except PARAFAC2xN was initialized with a "best out of 10" (positive) random starts method. This was necessary to avoid an excessive number of solutions that converged to local minima.[42] Each algorithm was fitted multiple times starting

from different random values. In total 50 converged models per algorithm and data set were compared to get an estimate of the precision and stability of the different algorithms.

MATLAB version R2022b (The MathWorks, Natick, MA USA) has been used for implementing the shift-invariant multi-linearity algorithm and the MCR-ALS routines. The PARAFAC2×N algorithm was taken from the GitHub repository: https://github.com/mdarmstr/parafac2x2, 24.03.2024. Area of feasible solution calculation were executed using the FACPAC software version 2.0 (http://www.math.uni-rostock.de/facpack/Downloads.html#Current_release, 24.03.2024)[46].

3.2 Simulated data

Data has been simulated to mimic a GC × GC-TOFMS data set with well-defined SNR using the function developed by Sorochan Armstrong et al[13]. The source code of the function is available under the URL https://github.com/mdarmstr/parafac2x2, 05.03.2024. Specifically, seven data sets emulating repeated measurements of two analytes eluting in retention-window (${}^1t_R \times {}^2t_R$) were simulated under different SNR conditions. The number of samples in each data set was set to 10 and the number of modulations in ${}^1D$ was set to 20 and the number of scans in ${}^2D$ was set to 200. The mass axis was clipped to a range of 761 *m/z* values. To study the performance of the algorithm in different SNR regimes, the SNR was varied from 3 to 0.025. Figure 4 shows examples of simulated measurements of the data sets with moderate SNR and with low SNR. While for the SNR of 0.1 peak shapes are visible in the contour plots of the 2D-TIC and the 2D-BPCs, at the low SNR hardly any peak-like structure is recognizable.

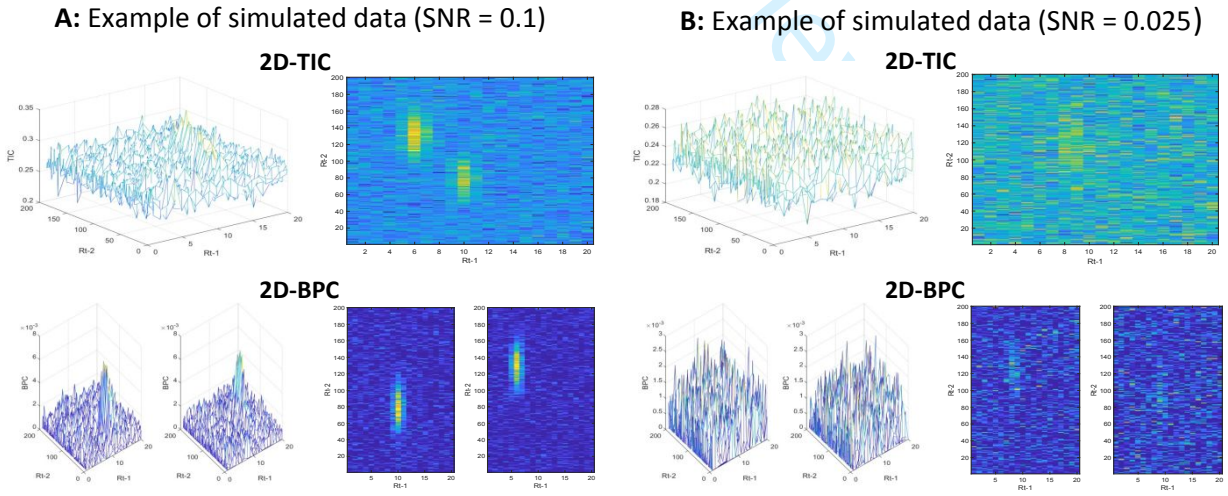


*Figure 4: **A:** Example of one simulated measurement at SNR of 0.1. **B:** Example of one simulated measurement at SNR of 0.025. The total ion chromatogram and the base peak chromatogram are shown to illustrate the noisiness of the simulated data.*

3.3 Experimental data

Previously published GC × GC-TOFMS data from a calibration experiment was used to compare the algorithm performance on real data. The two analytes present in the modeled ($^1t_R \times {}^2t_R$)-frame are the derivatized forms of salicylic acid and adipic acid (both molecules are derivatized with two TMS groups each, to cover the acid and hydroxy functionalities). For details on the derivatization procedure, we refer the reader to the original article.[13]

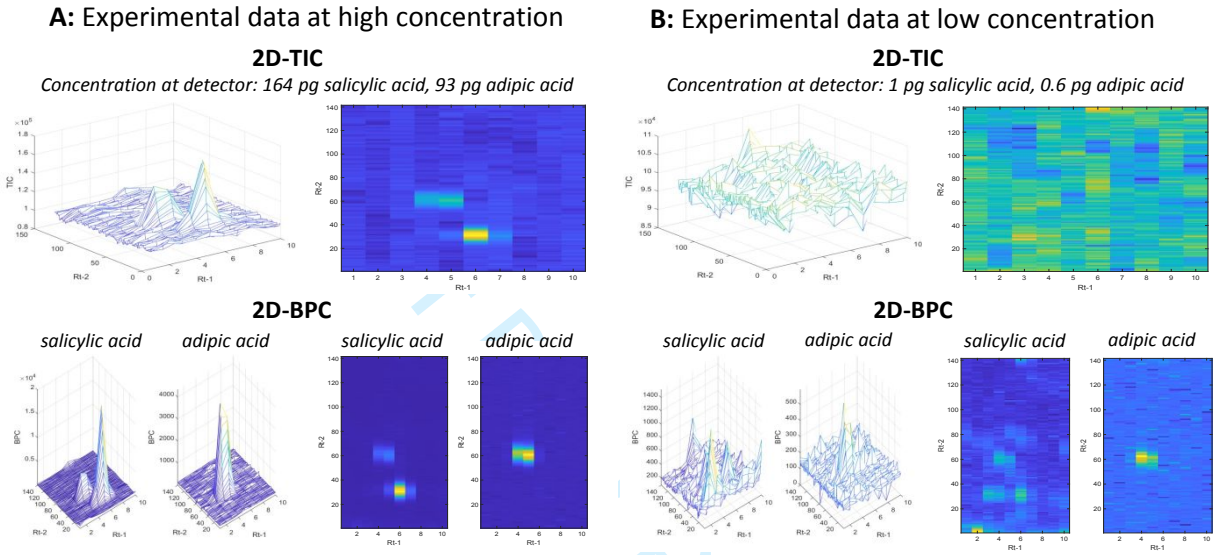


*Figure 5: **A:** Example of one of the triplicates at the highest calibration concentration considered (calibration point 6 in Table 1). The total ion chromatogram and the base peak chromatogram for salicylic and adipic acid are shown. The base peak chromatogram of salicylic acid is a fragment shared by adipic acid. **B:** Example of one of the triplicates at a lower calibration concentration (calibration point 12 in Table 1). The total ion chromatogram and the base peak chromatograms show the noisiness of the data at the low concentration. The noise is different from the simulated case as it has a lower frequency and is more structured.*

The original data set consists of 14 calibration points covering a range of injected analyte amount of 0.1 pg – 16,393 pg for salicylic acid and 0.1 – 9,311 pg for adipic acid on column. Each calibration point has been measured in triplicates. Since we were mostly interested in investigating algorithm performance in the low SNR domain, we discarded the first five calibration points from the data set to begin with and then continuously removed further calibration points to study the breakdown point for each algorithm. The highest calibration point in the first data set is 163.9 pg salicylic acid and 93.1 pg adipic acid from the analyte mass that reaches the instrument, and the highest calibration point in the last data set is 1 pg salicylic acid and 0.6 pg adipic acid.

Table 1: List of the calibration standards published by Armstrong et al. with their respective concentrations on column in pg. Standards 1-5 have been removed from the set for this study and standard 6 has been used to determine the precision in the extrapolation experiment (see description in 3.4).

Cal. standard	Salicylic acid [pg]	Adipic acid [pg]	Replicates
1*	16392.6	9311.1	3
2*	8196.3	4655.6	3
3*	2732.1	1551.9	3
4*	1092.8	620.7	3
5*	327.9	186.2	3
6**	163.9	93.1	3
7	82.0	46.6	3
8	27.3	15.5	3
9	10.9	6.2	3
10	5.5	3.1	3
11	2.7	1.6	3
12	1.0	0.6	3
13	0.5	0.3	3
14	0.1	0.1	3

3.4 Metrics for evaluation

The performance of the algorithms was assessed using different quantitative metrics. The fit measured as variance explained ($VarExpl$) was used to evaluate how good the different methods can explain the data (Equation 4-6). In Equation 5-6, $x_{n,m}$ is the entry in the n^{th} row and m^{th} column of the original data matrix $\mathbf{X}$ $(IKL \times J)$ and $\hat{x}_{n,m}$ is the entry in the n^{th} row and m^{th} column of the matrix $\hat{\mathbf{X}}$ $(IKL \times J)$, reconstructed from $\mathbf{C}$ and $\mathbf{S^T}$.

Equation 4: $$VarExpl = \left(1 - \frac{SSE}{SST}\right) * 100$$

where

Equation 5: $$SSE = \sum_n^{IKL} \sum_m^J (x_{n,m} - \hat{x}_{n,m})^2$$

and

Equation 6: $$SST = \sum_n^{IKL} \sum_m^J (x_{n,m})^2$$

The cosine similarity (also known as Tucker congruence) was calculated according to Equation 7 to assess how well the estimated mass spectra correspond to the true mass spectra. In Equation 7, $\mathbf{s}_{r,\text{est}}$ is the estimated mass spectrum and $\mathbf{s}_{r,\text{ref}}$ is the reference mass spectrum.

Equation 7: $$cosine\ similarity = \frac{\mathbf{s}_{\text{r,est}}\mathbf{s}_{\text{r,ref}}^{\text{T}}}{||\mathbf{s}_{\text{r,est}}||_F^2 ||\mathbf{s}_{\text{r,ref}}||_F^2}$$

The quality of the estimated concentrations was assessed in two different ways for the simulated data and for the real data. For the simulated data, calibration curves were fitted between the estimated concentrations and the true concentrations. The R^2 value and the bias (offset) of the calibration curve were evaluated as they are commonly used figures of merit in quantitative chromatographic analysis.

For the real calibration data, an extrapolation experiment was performed, in which the concentration of the highest calibration point (163.9 pg salicylic acid and 93.1 pg adipic acid) was successively predicted with models which were built on the $(9 - p)$ lowest calibration points. The pooled, relative standard deviation was used to assess the quantitative precision of the different methods across the fitted models after removing $p = 1,\ldots,6$ calibration standards.

4. Results and Discussion

4.1 Simulated Data

The results of using the different models on the simulated data set indicate larger performance differences between MCR-ALS, SIML, and PARAFAC2×N, which are summarized in Figure 6 and Figure 7.

While MCR-ALS and SIML achieve nearly the same fit values on the simulated data sets, PARAFAC2×N fits the data on average significantly worse. Moreover, Figure 6A also shows that the fit values achieved with PARAFAC2×N are subject to larger variation across the 50 repeated fits, which indicates that some of the solutions converged to local minima.[42] The performance difference between PARAFAC2×N and the other models becomes more pronounced when comparing accuracies of the estimated mass spectra at different noise levels (Figure 6B). The cosine similarity between the estimated and the underlying true spectra declines rapidly for all algorithms when the SNR is lower than 0.5.

In direct comparison, at SNR of 0.1 SIML has cosine similarity (Figure 6B) and R^2 values (Figure 6C) close to one but for MCR the performance starts to degrade. The performance of SIML is slightly better than MCR at SNR of 0.05 and similar to MCR at SNR of 0.025. In contrast, SIML-DN performed well at all noise levels studied and provided high cosine similarity and R^2 values even at an SNR of 0.025.

With respect to the calibration curves it is however noticeable, that SIML with denoising has a larger bias than MCR-ALS and SIML without denoising at higher SNRs (0.5 - 3). This difference vanishes at lower SNR values (0.025 - 0.1), at which the bias of MCR-ALS, PARAFAC2×N and SIML without denoising becomes larger than the bias of SIML with denoising. The reason for this is that the models without denoising have increasing difficulties separating the baseline from the analyte signal, which can also be emphasized by comparing the ^{2}D elution profiles and mass spectra shown in Figure 7 at SNR of 0.1 and 0.025. Especially, the comparison of the elution profiles and mass spectra at a SNR of 0.025 highlight the stability of the SIML-DN method, because it still provides reasonable estimates whereas the estimates of the other methods can hardly be qualified as chemical information.

Although SIML appears to be more stable than MCR-ALS based on the results of the simulation study, the differences are not huge, supporting MCR-ALS as a strong benchmark.

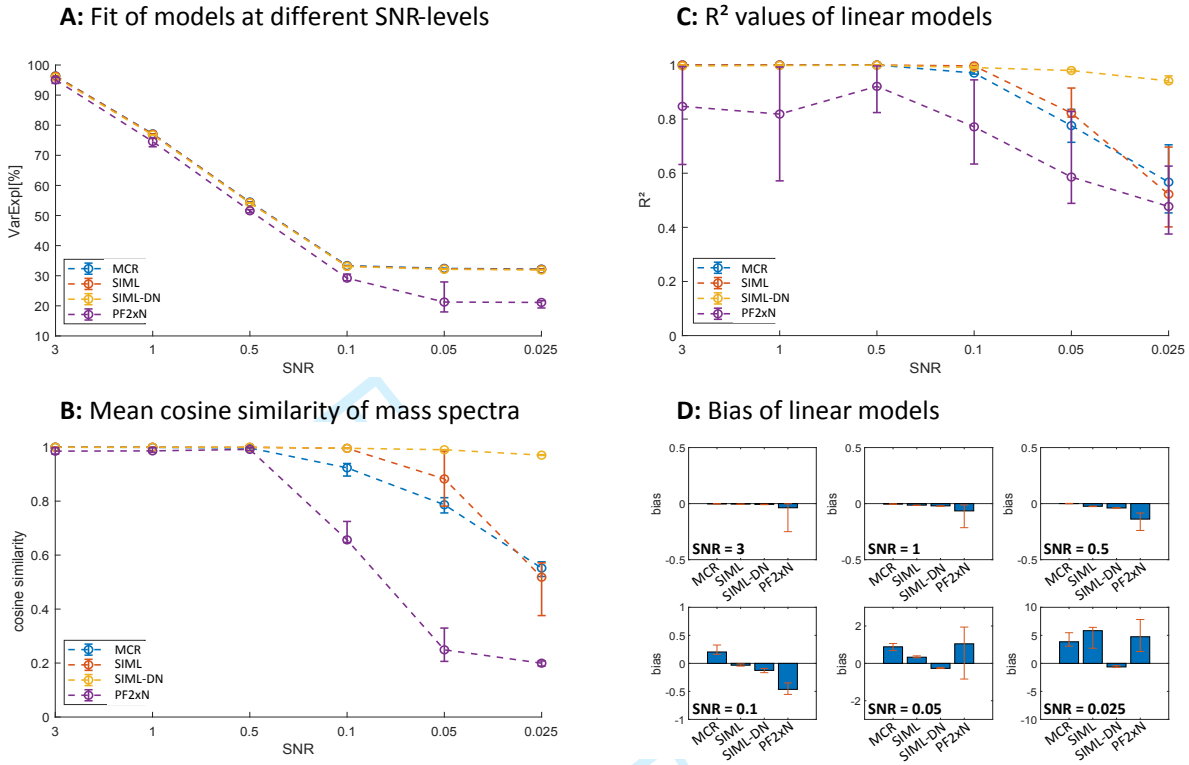


Figure 6: Summary of the performance of MCR, SIML, SIML-DN and PF2×N on simulated GC × GC-TOFMS data at different SNRs. In all plot are the mean values and standard deviations over 50 repeated fits at each SNR visualized. ***A:*** *variance explained indicating how well the different models describe the data. All models show the same trend that as the SNR descreases, the fit of the models gets worse. PF2xN shows significantly lower fit compared to all the other models.* ***B:*** *Mean cosine similarity of the estimated mass spectra and the true mass spectra. The models show distinct capability of modelling the true mass spectra at different SNR. The SIML-DN algorithm is most robust against high noise levels.* ***C:*** *R^2 values of linear models fitted on the peak areas and the known concentration values. The SIML-DN algorithm is most robust against high noise levels and PF2xN shows higher variability, probably due to the presence of local minima.* ***D:*** *Bias of linear models fitted on the peak areas and the known concentration values. The bias of MCR, SIML and PF2xN increases more with decreasing SNR than the bias of SIML-DN*

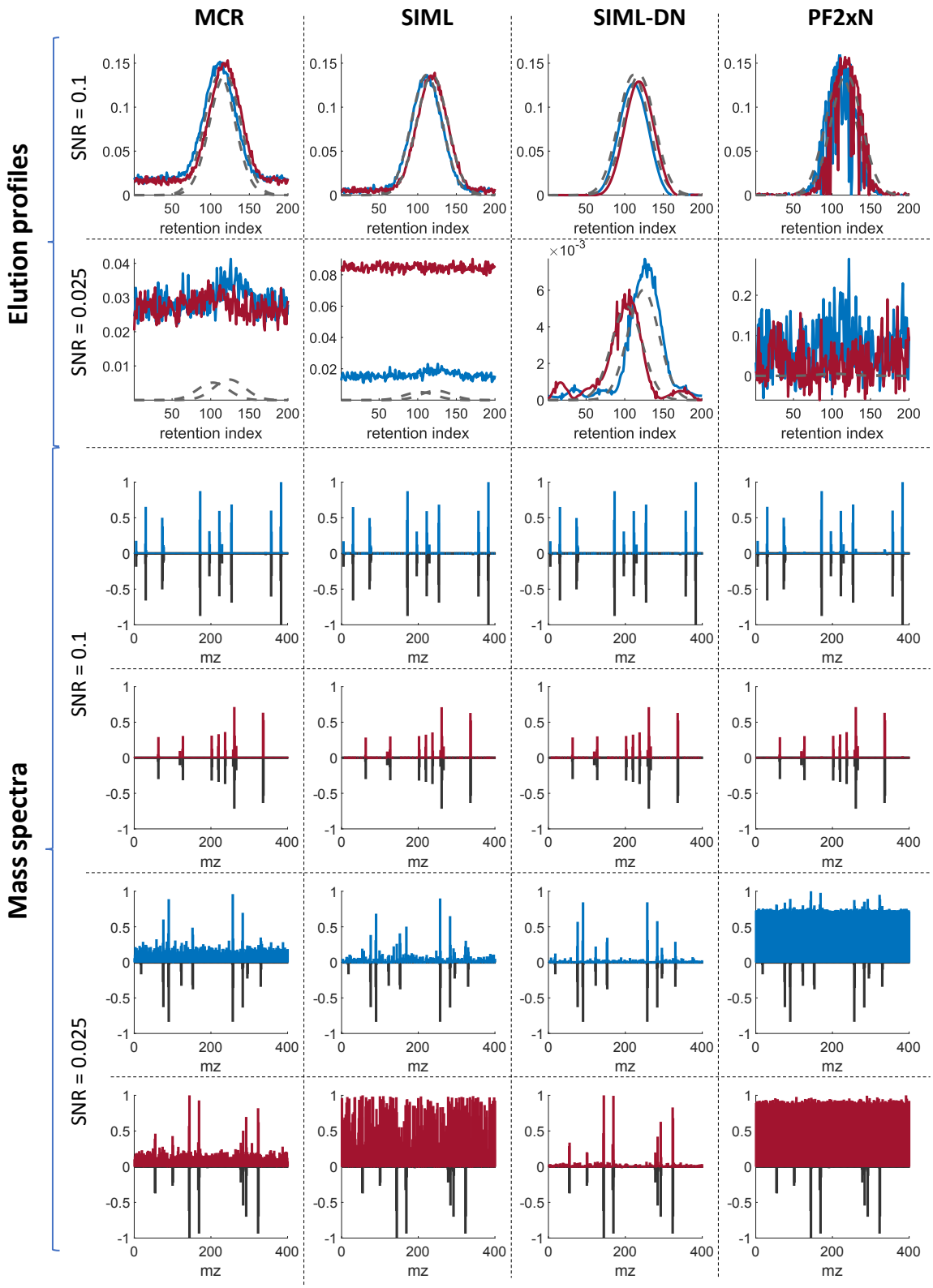


Figure 7: Comparison of estimated elution profiles and mass spectra at high and at low SNR. The colored elution profiles and spectra resemble estimates, while dashed and solid black lines show the true reference profiles and spectra.

4.2 Experimental Data

In comparison to the results for the simulated data, the situation changes quite dramatically when looking at the calibration data of salicylic and adipic acid. Although MCR-ALS fits the data better than SIML and SIML-DN (Figure 8A), the cosine similarities between the estimated and the true spectra are on average significantly worse than the estimates of SIML and SIML-DN (Figure 8B). Moreover, the rotational ambiguity of the MCR model translates to a larger variability in the estimated spectral profiles compared to SIML and SIML-DN. The same holds for the precision of the predictions of the calibration standard (163.9 pg salicylic acid / 93.1 pg adipic acid) shown in Figure 8C. The relative error is calculated from the predictions made with models built on the calibration data after removing 6, 7, … ,11 calibration points, as pointed out in section 3.4. In the extremes, a model was built on calibration data ranging from 0.1 to 1.1 pg and from 0.1 to 0.6 pg for salicylic acid and adipic acid, respectively, to predict concentrations of 163.9 pg salicylic acid and 93.1 pg adipic acid.

While the inter quartile range of the relative prediction error from the MCR models reaches from -0.5 to 0.1 (salicylic acid) and from -0.5 to 0.5 (adipic acid), the inter quartile range for the SIML and SIML-DN models reaches at most from -0.02 to 0.02 considering salicylic acid and adipic acid. The few high prediction errors from the SIML models for adipic acid can be related to the results from models built on calibration data after removing 11 standards. At this point the SIML models reach their breaking point (compare Figure 8B), while SIML-DN still provides reliable mass spectra and concentration estimates.

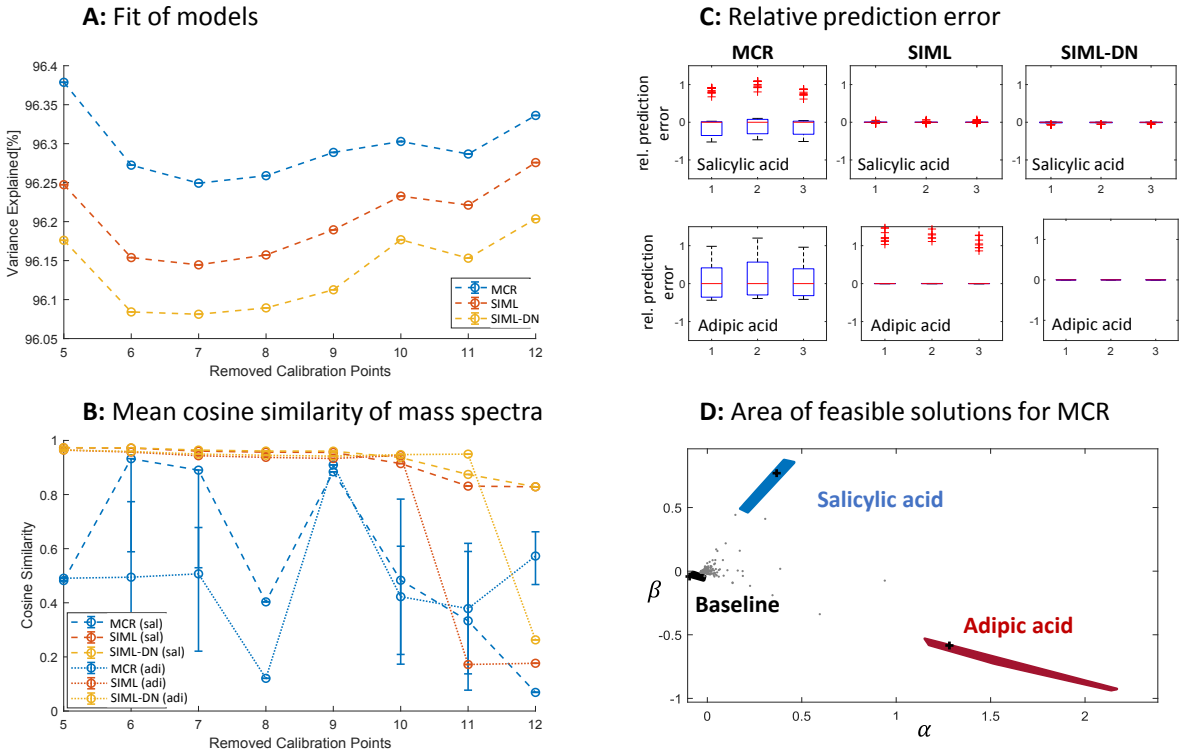


Figure 8: Summary of the performance of MCR, SIML and SIML-DN on a real GC × GC-TOFMS calibration data set. In all plot are the mean values and standard deviations over 50 repeated fits on increasingly smaller subsets of the calibation data visualized. ***A:*** *variance explained indicating how well the different models describe the data. All models show the same trend, however MCR tends to have a silghtly better fit, followed by SIML and SIML-DN which have the lowest fit. This trend follows the intuition that the least constrained model should have the highest fit.* ***B:*** *Mean cosine similarity of the estimated mass spectra and the true mass spectra. The estimates obtained from MCR show large variance which can be attributed to rotational ambuigity. The estimates of SIML and SIML-DN show high accuracy up to the removal of 11 and 12 calibration points, respectively.* ***C:*** *Relative prediction error that is made when the highest calibration standard is predicted with models trained on subsets of the calibration data set after removing up to 11 calibration points. The relative prediction error made with the MCR models is in the order of magnitude of ± 50 % while the relative prediction error for SIML and SIML-DN is < ± 2 % up to the point when 11 calibration points are removed. At this point the prediction error of the SIML model increases drastically for adipic acid.* ***D:*** *Area of feasible solution for MCR on the whole concatenated calibration data set. The area of feasible solution shows that MCR suffers substantially from rotational ambiguity, even on the data set including all calibration points.*

This big difference in the results is because MCR is suffering from rotational ambiguity (see Figure 8D) whereas SIML and SIML-DN provide unique solutions. Compared to the simulated data, the calibration data set is a harder challenge because the mass spectra as well as the concentration profiles are correlated. The correlation in the mass spectra can be traced back to shared fragments that stem from the TMS derivatization (specifically fragments 73,74, and 147).[47] Derivatization with TMS or other reagents is a common practice in gas chromatography and may cause similar problems for MCR because the mass spectra will contain many abundant, and common fragments. The correlation in the concentration profiles

is straightforward to explain because the calibration data represents a dilution series. In Figure 9 are the ^{2}D elution profiles and mass spectra shown from models built on calibration data after removing 6 and 11 calibration standards, respectively. In all cases the models that achieved the highest cosine similarity with their spectral estimates were selected for visualization in Figure 9. Among these models, the difference between the MCR and the SIML estimates is not very pronounced but still visible in the offset of the MCR elution profiles at the higher concentration level. The visualization of the estimates from the smallest calibration set shows once more that the denoising implemented in SIML-DN pays off because it allows for the extraction of useful chemical information under really challenging conditions.

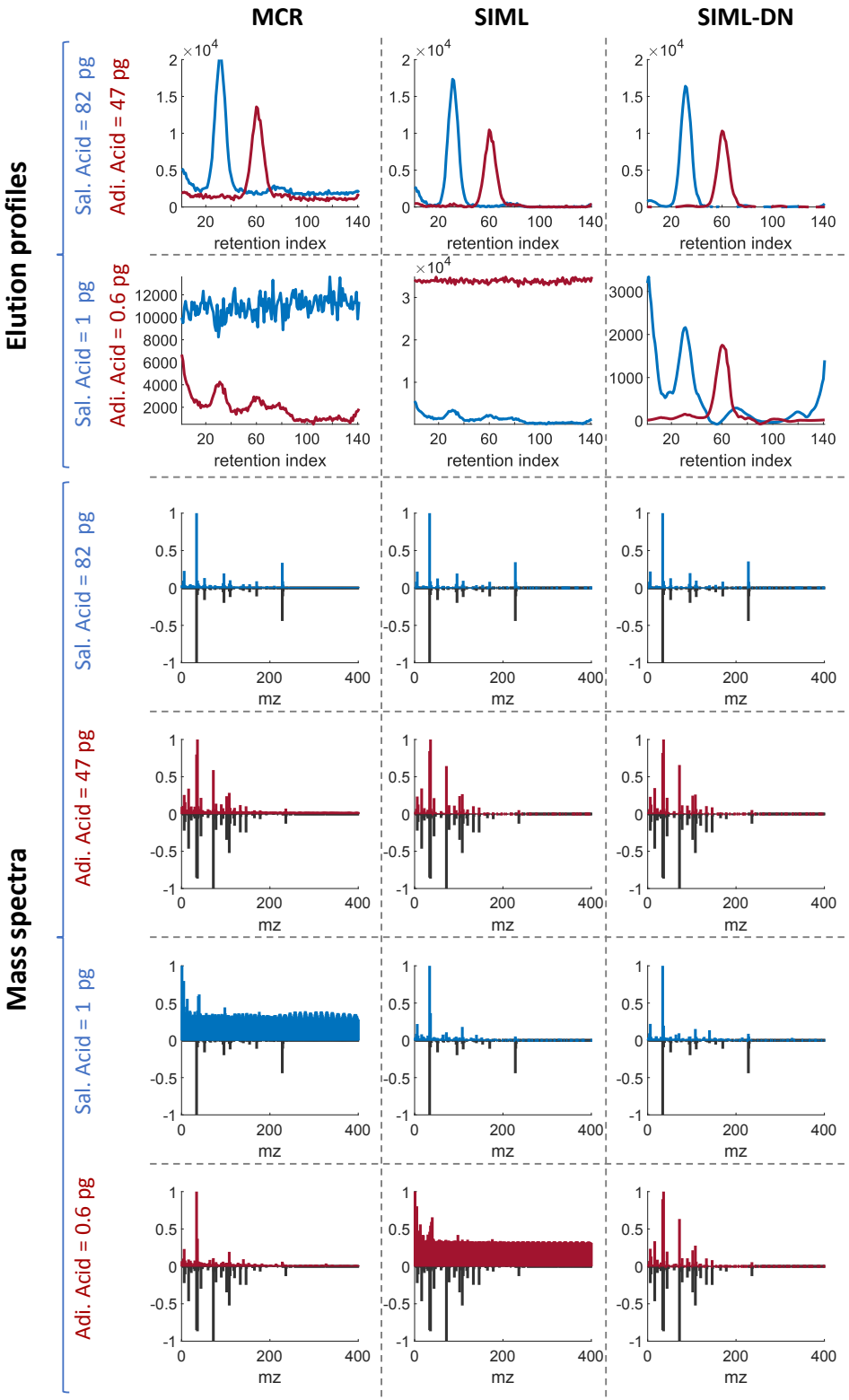


Figure 9: Comparison of estimated elution profiles and mass spectra at high and at low SNR. The colored elution profiles and spectra resemble estimates, while dashed and solid black lines show the true reference profiles and spectra.

5. Conclusions

In conclusion, the proposed shift-invariant multi-linearity (SIML) algorithm demonstrates a significant advancement in the analysis of comprehensive GC × GC-TOFMS data. By incorporating wavelet-based smoothing and Fourier-Transform based de-shifting into the multivariate curve resolution-alternating least squares (MCR-ALS) routine, the SIML algorithm successfully addresses the challenges of retention time shifts and high noise levels. Benchmarking against non-negativity constrained MCR-ALS and PARAFAC2×N methods reveals that SIML provides unique solutions and exhibits unmatched robustness against high noise levels, achieving impressive performance in both simulated and real data scenarios. The robustness is especially evident in lower signal-to-noise ratio (SNR) regimes, where SIML-DN maintains high accuracy in estimating mass spectra and concentrations. This enhances the reliability of compound identification and quantitative analyses in complex matrices, which is crucial for advancing the applications of GC × GC-TOFMS in various fields such as environmental science, food chemistry, and biological research.

6. Acknowledgements

The authors would like to express their gratitude to BASF SE for funding the PhD project that formed the basis of this research. Michael Sorochan Armstrong is funded by an MSCA grant (number 101106986).

7. References

1. Muscalu, A. M. & Górecki, T. Comprehensive two-dimensional gas chromatography in environmental analysis. *TrAC - Trends in Analytical Chemistry* vol. 106 225–245 Preprint at https://doi.org/10.1016/j.trac.2018.07.001 (2018).
2. Yu, M., Yang, P., Song, H. & Guan, X. Research progress in comprehensive two-dimensional gas chromatography-mass spectrometry and its combination with olfactometry systems in the flavor analysis field. *Journal of Food Composition and Analysis* vol. 114 Preprint at https://doi.org/10.1016/j.jfca.2022.104790 (2022).
3. Thong, A., Basri, N. & Chew, W. Comparison of untargeted gas chromatography-mass spectrometry analysis algorithms with implications to the interpretation and putative identification of volatile aroma compositions. *J Chromatogr A* **1713**, (2024).
4. Zanella, D. *et al.* The contribution of high-resolution GC separations in plastic recycling research. *Analytical and Bioanalytical Chemistry* vol. 415 2343–2355 Preprint at https://doi.org/10.1007/s00216-023-04519-8 (2023).
5. Stilo, F., Bicchi, C., Reichenbach, S. E. & Cordero, C. Comprehensive two-dimensional gas chromatography as a boosting technology in food-omic investigations. *Journal of Separation Science* vol. 44 1592–1611 Preprint at https://doi.org/10.1002/jssc.202100017 (2021).
6. Reichenbach, S. E., Tao, Q., Cordero, C. & Bicchi, C. A data-challenge case study of analyte detection and identification with comprehensive two-dimensional gas chromatography with mass spectrometry (GC×GC-MS). *Separations* **6**, (2019).
7. Prebihalo, S. E. *et al.* Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications. *Analytical Chemistry* vol. 90 505–532 Preprint at https://doi.org/10.1021/acs.analchem.7b04226 (2018).
8. Prebihalo, S. E., Reaser, B. C. & Gough, D. V. Multidimensional Gas Chromatography: Benefits and Considerations for Current and Prospective Users. *LCGC North America* 508–513 (2022) doi:10.56530/lcgc.na.zi3478f2.
9. Weggler, B. A. *et al.* A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software. *J Chromatogr A* **1635**, (2021).
10. Stefanuto, P. H., Smolinska, A. & Focant, J. F. Advanced chemometric and data handling tools for GC×GC-TOF-MS: Application of chemometrics and related advanced data handling in chemical separations. *TrAC - Trends in Analytical Chemistry* vol. 139 Preprint at https://doi.org/10.1016/j.trac.2021.116251 (2021).
11. Furbo, S., Hansen, A. B., Skov, T. & Christensen, J. H. Pixel-Based Analysis of Comprehensive Two-Dimensional Gas Chromatograms (Color Plots) of Petroleum: A Tutorial. *Anal Chem* **86**, 7160–7170 (2014).
12. Sudol, P. E., Ochoa, G. S. & Synovec, R. E. Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *J Chromatogr A* **1644**, 462092 (2021).
13. Sorochan Armstrong, M. D., Hinrich, J. L., de la Mata, A. P. & Harynuk, J. J. PARAFAC2×N: Coupled decomposition of multi-modal data with drift in N modes. *Anal Chim Acta* **1249**, (2023).
14. Zhang, Z., Ma, P. & Lu, H. Two-Way Data Analysis: Multivariate Curve Resolution: Noniterative Resolution Methods. in *Comprehensive Chemometrics* 137–152 (Elsevier, 2020). doi:10.1016/B978-0-12-409547-2.14875-9.
15. Lawton, W. H. & Sylvestre, E. A. Self Modeling Curve Resolution. *Technometrics* **13**, 617 (1971).
16. de Juan, A., Rutan, S. C. & Tauler, R. Two-Way Data Analysis: Multivariate Curve Resolution – Iterative Resolution Methods. in *Comprehensive Chemometrics* 325–344 (Elsevier, 2009). doi:10.1016/B978-044452701-1.00050-8.

17. Lin, C.-J. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Comput* **19**, 2756–2779 (2007).
18. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126 (1994).
19. Paatero, P. The Multilinear Engine: A Table-Driven, Least Squares Program for Solving Multilinear Problems, including the n-Way Parallel Factor Analysis Model. *Journal of Computational and Graphical Statistics* **8**, 854 (1999).
20. Parastar, H., Radović, J. R., Bayona, J. M. & Tauler, R. Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares ABC Highlights: Authored by Rising Stars and Top Experts. *Anal Bioanal Chem* **405**, 6235–6249 (2013).
21. Hoggard, J. C. & Synovec, R. E. Parallel factor analysis (PARAFAC) of target analytes in GC × GC-TOFMS data: Automated selection of a model with an appropriate number of factors. *Anal Chem* **79**, 1611–1619 (2007).
22. Sawall, M., Schröder, H., Meinhardt, D. & Neymeyr, K. On the Ambiguity Underlying Multivariate Curve Resolution Methods. in *Comprehensive Chemometrics* 199–231 (Elsevier, 2020). doi:10.1016/B978-0-12-409547-2.14582-2.
23. Tauler, R., Smilde, A. & Kowalski, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J Chemom* **9**, 31–58 (1995).
24. Olivieri, A. C. A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution − a tutorial. *Anal Chim Acta* **1156**, 338206 (2021).
25. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171 (1997).
26. Bro, R., Andersson, C. A. & Kiers, H. A. L. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. *J Chemom* **13**, 295–309 (1999).
27. Tauler, R. Multivariate curve resolution of multiway data using the multilinearity constraint. *J Chemom* **35**, (2021).
28. Schneide, P. A., Bro, R. & Gallagher, N. B. Shift-invariant tri-linearity—A new model for resolving untargeted gas chromatography coupled mass spectrometry data. *J Chemom* **37**, (2023).
29. Van Benthem, M. H. & Keenan, M. R. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *J Chemom* **18**, 441–450 (2004).
30. Jaumot, J. & Tauler, R. MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution. *Chemometrics and Intelligent Laboratory Systems* **103**, 96–107 (2010).
31. Olivieri, A. C., Neymeyr, K., Sawall, M. & Tauler, R. How noise affects the band boundaries in multivariate curve resolution. *Chemometrics and Intelligent Laboratory Systems* **220**, 104472 (2022).
32. Bro, R. & Sidiropoulos, N. D. Least squares algorithms under unimodality and non-negativity constraints. *J Chemom* **12**, 223–247 (1998).
33. De Juan, A. & Tauler, R. Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. in *Analytica Chimica Acta* vol. 500 195–210 (Elsevier, 2003).
34. Tauler, R., Smilde, A. & Kowalski, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J Chemom* **9**, 31–58 (1995).
35. Harshman, R. A. & Lundy, M. E. The PARAFAC model for three-way factor analysis and multidimensional scaling. in *Research methods for multimode data analysis* (eds. Law, H. G., Snyder Jr, C. W., Hattie, J. A. & McDonald, R. P.) 122–215 (Praeger, New York, 1984).
36. Carroll, J. D. & Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 283–319 (1970).

37. Cohen Jeremy E. and Bro, R. Nonnegative PARAFAC2: A Flexible Coupling Approach. in *Latent Variable Analysis and Signal Separation* (ed. Deville Yannick and Gannot, S. and M. R. and P. M. D. and W. D.) 89–98 (Springer International Publishing, Cham, 2018).
38. Zhang, X. & Tauler, R. Flexible Implementation of the Trilinearity Constraint in Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) of Chromatographic and Other Type of Data. *Molecules* **27**, 2338 (2022).
39. Cohen Jeremy E. and Bro, R. Nonnegative PARAFAC2: A Flexible Coupling Approach. in *Latent Variable Analysis and Signal Separation* (ed. Deville Yannick and Gannot, S. and M. R. and P. M. D. and W. D.) 89–98 (Springer International Publishing, Cham, 2018).
40. Kiers, H. A. L., ten Berge, J. M. F. & Bro, R. PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *J Chemom* **13**, 275–294 (1999).
41. Gallagher, N. B. Classical least squares for detection and classification. in 231–246 (2019). doi:10.1016/B978-0-444-63977-6.00011-0.
42. Yu, H. & Bro, R. PARAFAC2 and local minima. *Chemometrics and Intelligent Laboratory Systems* **219**, 104446 (2021).
43. Nardecchia, A. & Duponchel, L. Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging. *Talanta* **217**, 121024 (2020).
44. Fourier and Wavelet Transforms. in *Data-Driven Science and Engineering* 47–83 (Cambridge University Press, 2019). doi:10.1017/9781108380690.003.
45. Cooley, J. W. & Tukey, J. W. An algorithm for the machine calculation of complex Fourier series. *Math Comput* **19**, 297–301 (1965).
46. Sawall, M. & Neymeyr, K. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: Theoretical foundation, inverse polygon inflation, and *FAC-PACK* implementation. *J Chemom* **28**, 633–644 (2014).
47. DeJongh, D. C. *et al.* Analysis of trimethylsilyl derivatives of carbohydrates by gas chromatography and mass spectrometry. *J Am Chem Soc* **91**, 1728–1740 (1969).

1 Limitations of PARAFAC2 for GC-MS, GC×GC-TOFMS Data

1.1 The direct fitting method for PARAFAC2

It is easy to show that the PARAFAC2 constraint implies that elution profile loadings can change from sample-to-sample up to a rotation. One practical implication of this is that when overlapping peaks shift equidistantly then the PARAFAC2 constraint is not "violated". Hence PARAFAC2 is invariant to equidistant shifts. On the other hand if overlapping peaks shift differently, then PARAFAC2 can only approximate this and hence when the shifts are large for each peak relative to one another, it is to be expected that PARAFAC2 may run into problems. Note that the equal cross-product constraint allows for certain shape changes in addition to shifts.

To illustrate this, consider one-dimensional GC-MS data, for *M* scans, *N* mass spectra and *K* samples. A direct fit for a PARAFAC2 model is performed through the following decomposition:

$$X_k = P_k F D_k A^T \tag{1}$$

Where P_kF is an $M \times R$ matrix that describes the scores along the mode that is allowed to drift in or change shape. The assumption for the direct fitting model is that $\sum_k F_k P_k^T P_k F_k = \sum_k F_k^T F_k$ is comparable via the summation term for all *K* samples. Mathematically, this is explicit by solving for *F* in Equation 1:

$$F = \sum_{k=1}^{K} P_k^T X A D_k (D_k^T A^T A D_k)^{-1} \tag{2}$$

Where the proportion of variance explained by the orthonormal basis P_k is the sum of several such terms over $P_k^T X_K$ in which the relative variance explained by each *R* factor is assumed to be constant. However this assumption is not always fulfilled depending on how each chemical component drifts relative to each other. The orthonormal scores in P_k describe different variation within the nominal chemical space as shown in Figure 1 depending on their position relative to each other. This affects the calculation of the *F* matrix for the k^{th} slice. This was mentioned by Sorochan Armstrong [4].

1.2 Flexible Coupling PARAFAC2

The flexible coupling constraint relaxes this assumption somewhat, to allow for some drift between the calculated scores, B_k and what is termed the "coupling matrix", F in the additional term regularized via a constant μ_k [1]:

$$X_k = argmin_{B_k,P_k,F,D_k,A} \sum_{k=1}^{K} ||X_k - B_k D_k A^T||_F^2 + \mu_k ||B_k - P_k F||_F^2 \quad (3)$$

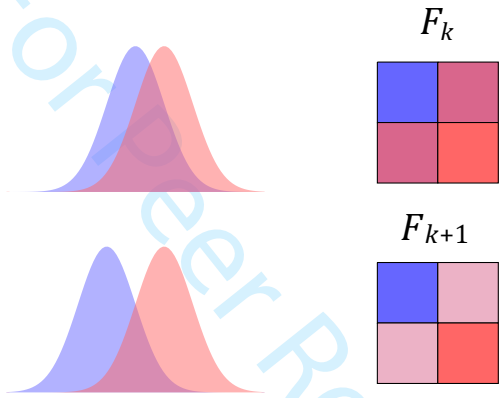


Figure 1: The effect of independent peak drift on the coupling matrix, F

1.3 Implication for multidimensional separations data

As a consequence of unfolding data from a GC×GC-TOFMS experiment as X = $X_I \in \Re^{I*K\times J\times L}$, drift along two modes is in effect drift along a single mode, and the sample mode L is equivalent to the sample mode K for one-dimensional experiments. Furthermore, drift along the first retention mode becomes far more pronounced along a combined retention mode, as even small changes in relative positions are distributed multiplicatively through the unfolding. This can be demonstrated using the Khatri-Rao, or column-wise Kronecker product for two series of elution scores: ${}^1T_l$ for the first dimensional scores, and ${}^2T_l$ for the second dimensional scores. The unfolded retention time, B_l in this case can be calculated if both elution modes are known:

$$B_l = {}^1T_l \otimes {}^2T_l \quad (4)$$

The relative change in the orthonormal basis P_I may contribute to modelling errors as F is calculated in the direct fitting PARAFAC2 model, as small independent changes in peak drift along the first mode manifest as large changes in the peak shape along the combined retention mode. For this reason, there are limitations to a PARAFAC2 model applied along a single, unfolded retention mode that is presumed to drift between samples.

2 SIML with integrated noise-filtering

2.1 VisuShrink with Soft-Thresholding for Wavelet Transform Denoising

The scores for SIML were de-noised at each iteration through a wavelet transform utilising soft thresholding, wherein the calculated wavelet coefficients are "shrunk" towards zero by a factor of their distance above the pre-specified threshold, before the inverse transform reconstructs the data according to the values of the modified coefficients. The threshold, λ was selected using the same methodology for a "hard" threshold (i.e. coefficients below this value are set to zero) using the criteria established by Donoho and Johnstone [3]:

$$\lambda = \sigma\sqrt{2log(M)} \tag{5}$$

where M is the length of the data, and σ is the estimated standard deviation of the additive noise component, following:

$$g(x_m) = f(x_m) + \sigma z_m \tag{6}$$

σ^2 is calculated using the wavelet coefficients at the highest frequencies as the median absolute deviation, $\left(\frac{MAD}{0.6745}\right)^2$. $z_m \sim N(0,1)$ is a vector of equal length to the observational data (x_m) containing white noise, $g(x_m)$ is the observational data and $f(x_m)$ is the true signal. The weights of the wavelet coefficients, w are modified as w' according to the soft thresholding method in[2]:

$$w' = \begin{cases} \text{sign}(w)(|w| - \lambda), & \text{if } |w| > \lambda, \\ 0 & \text{if } |w| \leq \lambda. \end{cases}$$

where coefficients below the threshold are set to zero, but those above the threshold are reduced in absolute magnitude by a factor corresponding to the threshold itself.

2.2 Noise filtering on estimated factors versus noise filtering of raw data

Two smoothing approaches are compared: Smoothing of the raw data, prior to modelling the data with SIML, and applying wavelet-based smoothing inside the ALS routine on the estimated elution profiles (SIML-DN). For the comparison, data sets with low SNR (0.05 and 0.025) have been simulated, as described in section 3.2 of the article. Smoothing of the raw data has been performed using the wavelet-based denoising procedure described in section 2.1 of the supporting information and by applying moving average filters of different window size (3, 5, 13). Since only white noise is added in the data simulation to adjust the SNR, moving average is the optimal filter in this case. [5]

Figure 2 shows the effect of different filters applied to the raw data. The blue lines indicate the results obtained for SIML (no additional smoothing inside the ALS routine) and the orange lines indicate the results obtained for SIML-DN (additional smoothing inside the ALS routine). The first case, "not smoothed" is used as benchmark (indicated as black, dashed line), for which no smoothing of the raw data has been performed. This is compared to the cases "wvlt smooting", "movmean 3", "movmean 5", and "movmean 13" for which the raw data has been smoothed prior to modelling with SIML / SIML-DN. As performance indicators the model fit, the spectral similarity and the linearity of the calibration curve (fitting modeled score values to the known, true concentrations) have been chosen.

The results depicted in Figure 2A-B show that smoothing of the raw data improves the model fit for SIML and SIML-DN significantly for the different SNR (0.05 and 0.025), compared to benchmark. Figure 2C-D show that the spectral similarity of SIML and SIML-DN is very different when no prior smoothing of the raw data is applied. The difference between the model results is reduced if smoothing of the raw data is applied, but the spectral similarities achieved with both models after smoothing the raw data do not exceed the benchmark. Figure 2E-F show how well the modeled scores resemble the known, true concentrations. Again, large differences can be observed if no smoothing of the raw data is applied. At SNR 0.05, higher R^2 values are achieved with both models (SIML / SIML-DN) compared to the benchmark, if the raw data is smoothed with a moving average filter. However, at SNR 0.025, no smoothing strategy provides higher R^2 values than the benchmark. The selected experimental conditions are very favorable for the moving average filter, because only stationary, white noise is added. This is an idealized situation opposed to more complicated noise characteristics found in real data.[6] Assuming a situation in which multiple measurements have been carried out under slightly varying conditions, it is likely that the inherent noise structure will vary from measurement to measurement. Therefore, the noise structure of a concatenated elution profile in the augmented matrix $\mathbf{X}$ (compare section 2.1 in the paper) wont be stationary under these

assumptions. Hence, we assume that the wavelet-based smoothing is a reasonable, yet certainly not perfect choice given the investigated data. Comprehensive parameter optimization of the level of resolution and the selected kernel function have not been carried out.

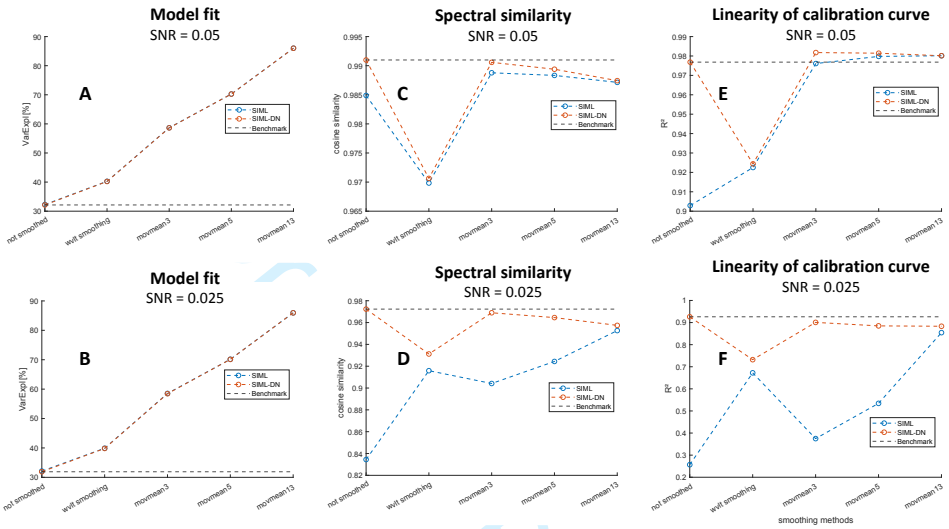


Figure 2: Comparison of different approaches for smoothing.

References

[1] Jeremy E Cohen and Rasmus Bro. Nonnegative parafac2: A flexible coupling approach. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 89–98. Springer, 2018.

[2] David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995.

[3] David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

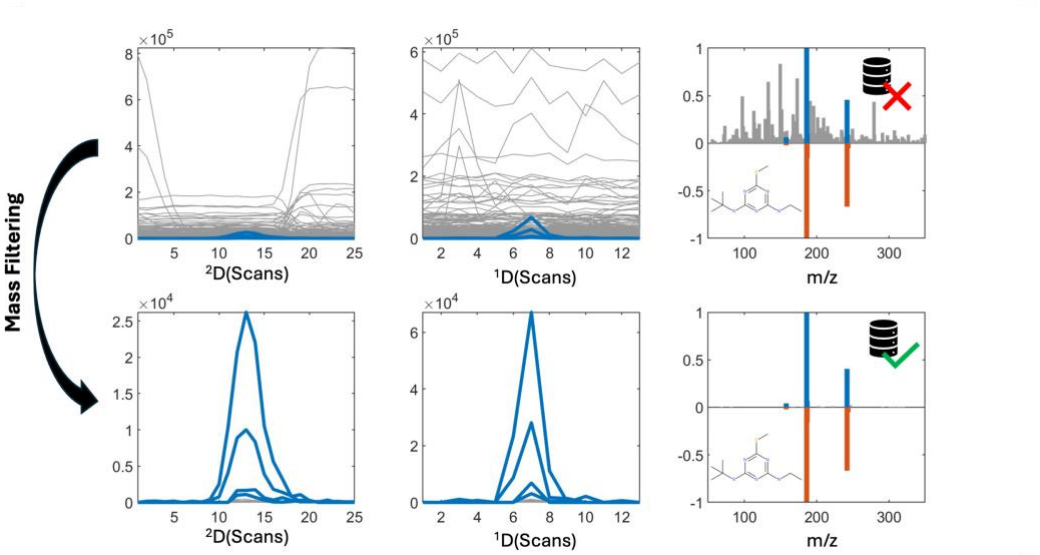
[4] Michael Sorochan Armstrong. *Decomposition and Feature Selection of Comprehensive 2-DimensionalGas Chromatography-Time-of-Flight Mass Spectrometry (GC× GC-TOFMS) Data*. PhD thesis, University of Alberta, 2021.

[5] Steven W. Smith. Digital Signal Processing – A practical guide for Engineers and Scientists. p. 277-284 (2002)

[6] Alexander B Fialkov, Urs Steiner, Steven J. Lehotay, Aviv Amirav. Sensitivity and noise in GC-MS: Achieving low limits of detection for difficult analytes. 260, 31-48, 2007

Paper 4

Schneide P-A, Kronik OM. A signal processing workflow for suspect screening in LC×LC-HRMS: Efficient extraction of pure mass spectra for identification of suspects in complex samples using a mass filtering algorithm. Analytical Chemistry (submitted)



A signal processing workflow for suspect screening in LC×LC-HRMS: Efficient extraction of pure mass spectra for identification of suspects in complex samples using a mass filtering algorithm

Journal:	*Analytical Chemistry*
Manuscript ID	ac-2024-042882
Manuscript Type:	Article
Date Submitted by the Author:	12-Aug-2024
Complete List of Authors:	Schneide, Paul-Albert; University of Copenhagen Faculty of Life Sciences, Food Science; BASF SE, Analytical Science Department Kronik, Oskar; University of Copenhagen Faculty of Life Sciences, Plant and Environmental Sciences

A signal processing workflow for suspect screening in LC×LC-HRMS: Efficient extraction of pure mass spectra for identification of suspects in complex samples using a mass filtering algorithm

Abstract

In contrast to the large number of liquid chromatographic high-resolution mass spectrometry (LC-HRMS) data processing workflows made for suspect and non-target screening. The workflows for comprehensive two-dimensional LC (LC×LC)-HRMS workflows are limited. In this work, we propose a two-step workflow to extract pure mass spectra in which a mass filtering (MF) algorithm is implemented that is grouping extracted ion chromatograms (EIC) based on their similarity in the first (^{1}D) and second dimension (^{2}D), followed by multivariate curve-resolution (MCR) on the extracted EICs to address potential co-elution. The workflow is referred to as MF+MCR. The purity of the mass spectra of the proposed workflow was compared to extracting the mass spectra at peak apex (PAM), using the MF approach alone, or using MCR without prior mass filtering. The workflows were tested on a pulsed elution-LC×LC-HRMS data from a wastewater effluent extract. This study showed that the proposed workflow correctly identified 25 suspect compounds compared to MF, MCR, and PAM approaches where 23, 16, and 10 suspect compounds could be identified. The nine suspects which could not be identified using MCR compared to the MF+MCR all had low total signal contributions i.e. low intensities compared to the TIC. This showed that adequate preprocessing prior to MCR is advantageous for trace level analysis.

The proposed workflow proved to be a promising approach to extract pure and high-quality mass spectra for subsequent identification from LC×LC-HRMS.

1 Introduction

Broad screening methods such as non-target and suspect screening methods are becoming more widely used for monitoring of complex samples such as environmental samples[1–4]. In recent years, multiple open-source software tools and data analysis workflows have been published to facilitate the demanding job of signal processing and compound identification of the information rich chemical data obtained from liquid chromatography hyphenated to high-resolution mass spectrometry (LC-HRMS).[5–8] Using tandem mass spectrometry, it is possible to associate pure product ion spectra (MS2) to specific pre-cursors (MS1), which, in combination with the rapidly

growing open-source mass spectral data bases, is a huge ressource.[9] The access to data bases has also been greatly facilitated by tools like MASST[10] or MS-DIAL[7] which allow querying MS2 spectra against large, curated data bases and repositories. To leverage these new capabilities, it is, however, crucial to extract pure MS2 mass spectra from the raw data.[7] Data analysis workflows in many open-source tools follow essentially the same steps, although different algorithms may be used[11]: First, EICs are constructed using e.g., region of interest (ROI) binning[12], secondly, peak detection is performed and peak boundaries are determined, followed by multiple filtering steps to discard uninformative EICs and keep only the informative ones.[13] Subsequently, the found features are aligned and grouped together based on their similarity to form component mass spectra (componentization).[14,15] While feature detection workflows have proved as valuable tools for data analysis, it is often argued that they do not suffice to extract pure mass spectra in cases of chromatographic co-elution and suffer from producing false positive and false negative results.[13,16,17] To the best of our knowledge, current feature detection workflows including MS-DIAL 4.0, MZmine 3, and XCMS are not able to process comprehensive two-dimensional liquid chromatographic hyphenated the high-resolution mass spectrometry (LC×LC-HRMS) data utilizing its inherent data structure.[7,8,11,15,18,19]

In addition to feature detection based approaches, curve-resolution and tensor decomposition have been applied to LC×LC-MS, LC-HRMS, GC×GC-MS, and GC-MS.[20–28] Multivariate curve resolution (MCR) and parallel factor analysis 2 (PARAFAC2) are commonly used methods in chemometrics that aim to extract chemically meaningful, compressed representations of the original chromatographic data consisting of components with concentration profiles and ideally pure mass spectra.[23] The motivation for applying such methods is that chromatographically co-eluting peaks can be deconvoluted and baseline signals separated yielding accurate estimates of the true analyte mass spectra and concentration profiles. In most cases, MCR and PARAFAC2 are iteratively fitted on small regions of the chromatographic data set to reduce the computational complexity of the task.[22,29] A conceptual description of MCR and PARAFAC2 and their application to 1D-LC-MS and 2D-LC-MS is given in Figure 1. Detailed mathematical descriptions of the methods can be found in the supporting material S1. A discussion of their model structures is beyond the scope of this study, and we refer the reader to the large body of literature on these subjects.[20,30–34]

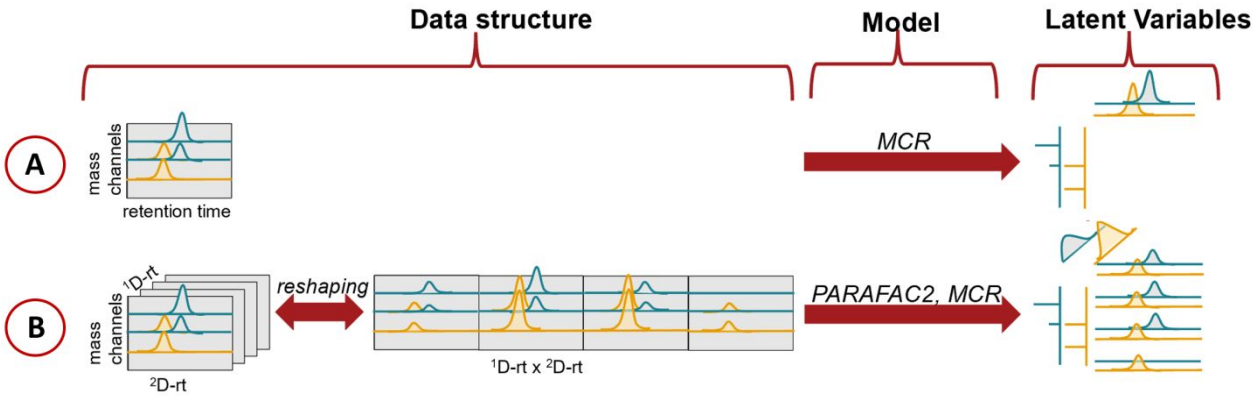


***Figure 1**: Description of the application of multivariate curve resolution and parallel factor analysis to different chromatographic data sets. Multivariate curve resolution and parallel factor analysis decompose windows of the original dataset into a set of latent variables i.e. components (low rank approximation). In all cases, the components consist of estimates for mass spectra, elution profiles and relative concentrations. **A**: If only a single 1D-LC measurement exists, the resulting dataset can only be decomposed using multivariate curve resolution. The relative concentrations of the analytes can be obtained by normalizing the elution profile estimates. **B**: In the case of LC×LC-HRMS, every sample has a data structure that matches the PARAFAC / PARAFAC 2 model. In this case, the ^{1}D is modeled analogous to the samples in 1D-LC and the ^{2}D is modeled analogous to the normal retention time axis, allowing profiles to be shifted.*

Due to the high number of chemical compounds present in the samples typically investigated in suspect and non-target screening, co-elution in LC-HRMS is a challenge, which renders the mass spectra to be a mix of multiple compounds. This increases the risk of wrongly identifying compounds as the mass spectra used for the spectral library search could contain fragment and precursor ions from more than one compound.[10] As a means of minimizing co-elution and impure mass spectra, LC×LC has been used due to its superior peak capacity compared to LC-HRMS[35–37]. In contrast to LC-HRMS data where relatively mature data processing workflows have been published, the same cannot be said for LC×LC-HRMS. Both, proprietary and open-source software packages exist for LC×LC data, but their focus has not been on signal processing so much as data visualization and method development.[38,39] Peak detection and extraction of pure mass spectra is therefore often laborious in LC×LC-HRMS. Navarro-Reig *et. al.,*[21] proposed a data processing workflow based on MCR where they showed that they could extract the most intense fragments of the lipid monogalactosyldiacylglycerol. However, for analysis of micro-pollutants present at low levels in environmental samples, it has been described that MCR is challenged in detecting compounds with low signal-to-noise ratio. [23,40–42]

In this study, we present a signal processing workflow for extracting pure mass spectra from LC×LC-HRMS data to facilitate more reliable compound identification for suspect screening. The

mass filtering (MF) method proposed and developed in this study extracts pure mass spectra of low and very low abundant analytes, which is hypothesized to increase the success-rate in compound annotation using mass spectral data bases (compare Figure 2).

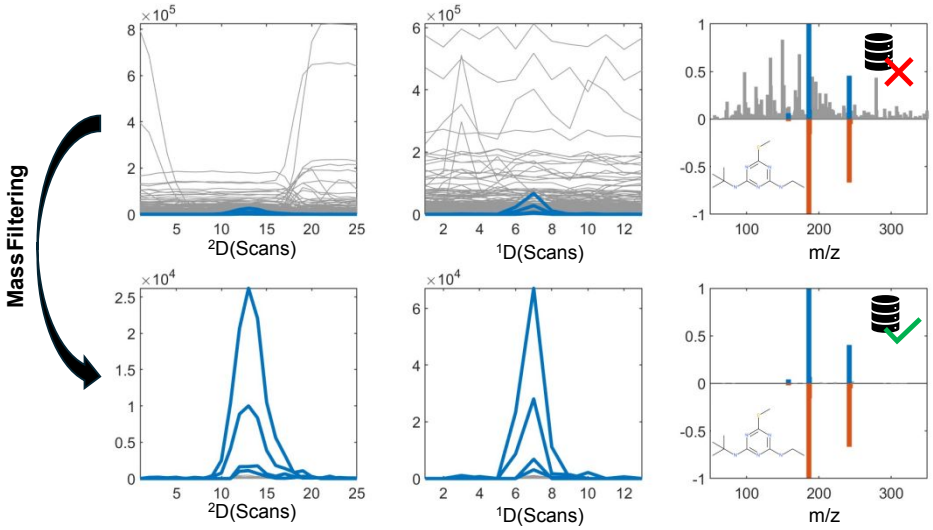


***Figure 2**: The result of applying mass filtering (MF) for the extraction of pure MS2 mass spectra is visualized. The upper row shows the 1D- and 2D-EICs and the mass spectrum of Terbutryn spectrum extracted from the peak apex. The lower row shows respectively the 1D and 2D-EICs and the mass spectrum of Terbutryn after applying MF. Querying the MF extracted mass spectrum against mass spectral libraries[10] yielded a successful hit confirming the presence of Terbutryn in the sample, whereas no hit was found using the mass spectrum extracted at peak apex.*

The performance of the developed workflow including the MF method to identify compounds is compared to using the mass spectra obtained from MCR and at peak apex (PAM), respectively. Specifically, the number of correctly grouped diagnostic ions and correctly identified compounds will be used to evaluate the performance of the three previously mentioned approaches. It was hypothesized that the presented workflow would improve the grouping of EICs for trace compounds compared to the MCR and PAM approaches due to its independence of peak intensity.

2 Materials and methods

2.1 Dataset

The dataset used in this study was the dataset previously published in Kronik et al[43] where the focus was on the method development and comparison of one dimensional and two-dimensional data. In this study, the main focus was on the signal processing of the data.

2.2 Chemicals and sample preparation

LC-MS grade acetonitrile (ACN) with 0.1% formic acid (FA) (v/v) and Methanol (MeOH) were bought from ChemSolute and Honey Well, respectively. LC-MS grade water was bought from ChemSolute, TH. Geyer. The FA and ammonium acetate (AmAc) for buffering the mobile phases were bought from Carlo Erba – Reagents and Merck, respectively.

A wastewater effluent sample was sampled from a wastewater treatment plant in Avedøre, Denmark and sample preparation was done according to the procedure described in Tisler et al[44]. The wastewater effluent sample was analyzed at a relative enrichment factor of 50.

2.3 Chromatographic conditions

The first dimension (^{1}D) of the LC×LC-HRMS platform was operated in pulsed elution mode, where pulses of organic modifier was sent through the ^{1}D column followed by a pause state where the organic modifier was decreased the 1 %B to decrease the eluotrophic strength of the mobile phase. This allowed for long second dimension (^{2}D) gradients. The pulse time was 0.9 min and pause time was 8.8 min giving a total modulation of 9.7 min. The ^{2}D gradient increased from 1-99 %B in 9 minutes followed by a 0.7 min equilibration step. Stationary phase assisted modulation was performed with two trap columns of UPLC VanGuard BEH C18 pre-columns (2.1 mm i.d. × 5 mm, 1.7 µm). Before the trapping the ^{1}D eluent was diluted with a 1.8 mL/min flow rate of water:ACN:FA (98.9:1:0.1% v/v). The mobile phase was 5 mM AmAc in water and MeOH in the ^{1}D and water and ACN with 0.1 %FA. An Acquity CSH PFP (2.1 i.d. × 50 mm, 1.7 µm, Waters), and an Acquity CSH C18 (2.1 i.d. × 30 mm, 1.7 µm, Waters) was used for the ^{1}D and ^{2}D, respectively. Both columns were thermostatted at 30 °C. All ultrahigh pressure liquid chromatography (UHPLC) modules were from Waters Corporation (Milford, United States) and controlled through Mass Lynx V4.1.

A Synapt G2-SI quadrupole time-of-flight (Q-TOF) was operated in positive mode using the MS^E setting where a low energy trace (MS1) and a high energy trace (MS2) is obtained in which more fragmentation occurs. The collision energies of MS1 and MS2 were 10 and ramping from 10-40 V, respectively. The scan time was 0.3 seconds. Lock mass correction was performed by infusing leucine enkaphaline every 30 sec. The remaining MS parameters can be seen in Tisler et al.[44]

2.4 Data Analysis Workflow

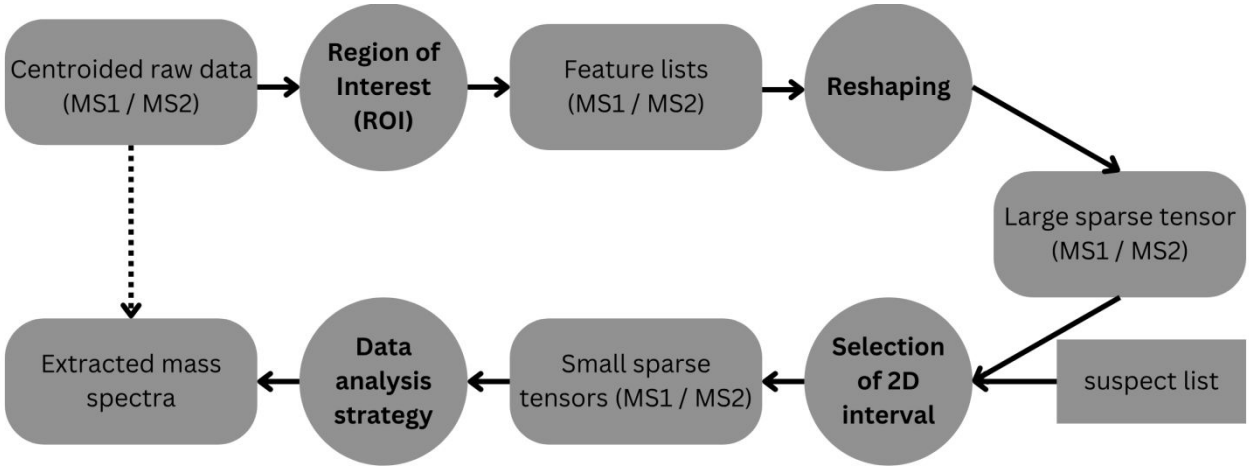


***Figure 3**: Visualization of the data analysis workflow.*

The centroided raw data (MS1 and MS2) was given as input to the data analysis workflow depicted in Figure 3 to extract purified mass spectra for every suspect as output. As first data pre-processing step region-of-interest (ROI) binning was performed to transform the centroided raw data into a feature list. The ROI procedure was parametrized setting the noise threshold to 0 (to avoid excluding low abundant peak fragments), the mass error to 16 ppm and the minimum peak width to 4 scans (i.e. 2.5 seconds). The feature lists were reshaped into a sparse tensor structure using the sparse tensor functions in the Tensor toolbox v.3.6 (www.tensortoolbox.org, 28.06.2023).[45] In the next step, EICs of the $[M+H]^+$ in MS1 of all suspects were used to create retention time windows in the 2D space (${}^1t_r \times {}^2t_r$) surrounding the peaks. This procedure was semi-automated using a 2D peak picking algorithm[46] and pre-defined interval width for ${}^1t_r$ and ${}^2t_r$ respectively. The interval boarders surrounding each peak were set to be ± 6 modulations in ${}^1t_r$ and ±13 scans in ${}^2t_r$, relative to the peak maximum. Based on the two-dimensional retention time window found using the described procedure, the large sparse tensors of the MS1 and MS2 data were cut into smaller tensors for each suspect. The reduced size tensors were used as input for the data analysis step.

Four different data analysis strategies were used to extract pure MS2 mass spectra for each suspect: 1) Peak Apex Method (PAM), 2) MCR, 3) Mass Filtering (MF) and 4) MF+MCR.

1) With PAM, mass spectra were extracted from the MS2 mass spectrum at the apex of the $[M+H]^+$ ion of the suspect found in MS1 within the selected ${}^1t_r \times {}^2t_r$ frame.
2) MCR decomposes the unfolded, reduced size tensor of the MS2 data into two bilinear component matrices, depicting the concatenated elution profiles and the mass spectra of all

compounds eluting in the given time frame. An alternating least squares algorithm was developed in-house applying non-negativity constraints for the estimation of both factor matrices. More details regarding rank analysis, initialization, and convergence can be found in the supporting material (S2). From the set of mass spectra, the one with the highest intensity of the $[M+H]^+$ ion was selected.

3) Mass filtering is a method presented herein that extracts EICs from MS2 based on their cosine similarity with the elution profiles of the $[M+H]^+$ ion in MS1. The similarities are calculated for each EIC along the ${}^1t_r \times {}^2t_r$ intervals that were used to cut out the smaller sized tensors (MS1 and MS2). A final similarity score is calculated as the product of the cosine similarities calculated along ${}^1t_r$ and ${}^2t_r$. Only EICs that achieved a final similarity score of > 0.7 were extracted. More details regarding the implementation can be found in the supporting material (S2).
4) Mass filtering at a lower threshold of 0.5 was used to remove noise prior to the deconvolution via MCR.

2.5 Compound identification

Database queries

The MS2 spectra of the respective compounds were queried to a database using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/, 04.03.2024) on the GNPS website (http://gnps.ucsd.edu, 04.03.2024).[10] Since the pipeline is built for MS data measured in data dependent acquisition, the $[M+H]^+$ peak was set as precursor ion and all EICs with larger m/z-ratio were discarded.

The data was filtered by removing all MS/MS product m/z within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The tolerance of precursor and product m/z was set to 0.01 Da. The library spectra were filtered like the input data.[10] Additionally, reference spectra from literature were collected and used for cross-checking the results from the MASST workflow.

Evaluation of suspect hits

The similarity between extracted and reference spectra and the number of shared fragments were used as quality metrics to compare the data analysis strategies. A successful confirmation of the

presence of an analyte was defined by the following two criteria (1) cosine similarity > 0.5 between extracted mass spectrum and reference mass spectrum, and (2) at least two shared fragments with the reference spectrum.

To assess the relationship between the intensity of the analyte signal and background signal quantitatively, the total signal contribution (TSC) was calculated for the different suspects. In Eq. (1), $SS_{suspect}$ is the sum of squares of the suspect signal and SS_{total} is the sum of squares of the total signal in a ${}^{1}t_r \times {}^{2}t_r$ frame. The range of TSC goes from zero to one. Values of the TSC metric close to zero indicate a small signal contribution of the suspect fragment ions to the total sum of squares and values close to one indicate a large signal contribution.

$$TSC = \frac{SS_{suspect}}{SS_{total}} \quad (1)$$

3 Results and Discussion

3.1 Extracting pure mass spectra from wastewater sample

In total 25 of the 42 suspects could be confirmed by at least one of the four data analysis strategies. All 25 confirmed suspects were correctly identified using MF+MCR, whereas MF, MCR, and PAM confirmed 23, 16, and 10 suspects, respectively (Table 1). The differences in performance could be explained by two factors: 1) the value of the *TSC*, and 2) the degree of co-elution or background noise on the EICs.

From the 25 identified suspects, the suspects with the 10 highest TSC were confirmed using MCR for extraction of mass spectra (Table 1). In contrast, only two of the 10 suspects with the lowest TSC could be confirmed using the MCR extracted spectra (Table 1) in the data base search. When the TSC was below 0.0013, MCR did not successfully extract satisfactory mass spectra (Table 1). This was due to the fact that the MCR algorithm converges towards a solution by minimizing the sum of squares i.e. when TSC, calculated according to Eq. (1), was too low a very high number of components would be required to describe the data satisfactory. Such a situation is exemplified in Figure 4**A**, in which the summed EICs belonging to 1-methylbenzotriazole are shown together with the TIC of all EICs. Note, that the TIC of the 1-methylbenzotriazole fragments was scaled by a factor of 10 for better visualization. Since the noise present in LC×LC-HRMS data is caused by many independent factors, we observed that extracting pure mass spectra by estimating a meaningful low-rank approximation of the data was not possible. In fact, even a higher model

complexity of the MCR model did only help to a limited degree to reduce the noisiness of the extracted mass spectra as can be seen in Figure 4**A** (blue colored m/z values belong to noise, orange-colored belong to suspect). Although the noisiness of the spectra could be reduced marginally going from a 10 component model to a 20 component model, the noisiness increases again for higher number of components due to numerical instability of the algorithm. This hampers the use of MCR for trace level analysis of e.g. environmental contaminants without further pre-processing of the data.

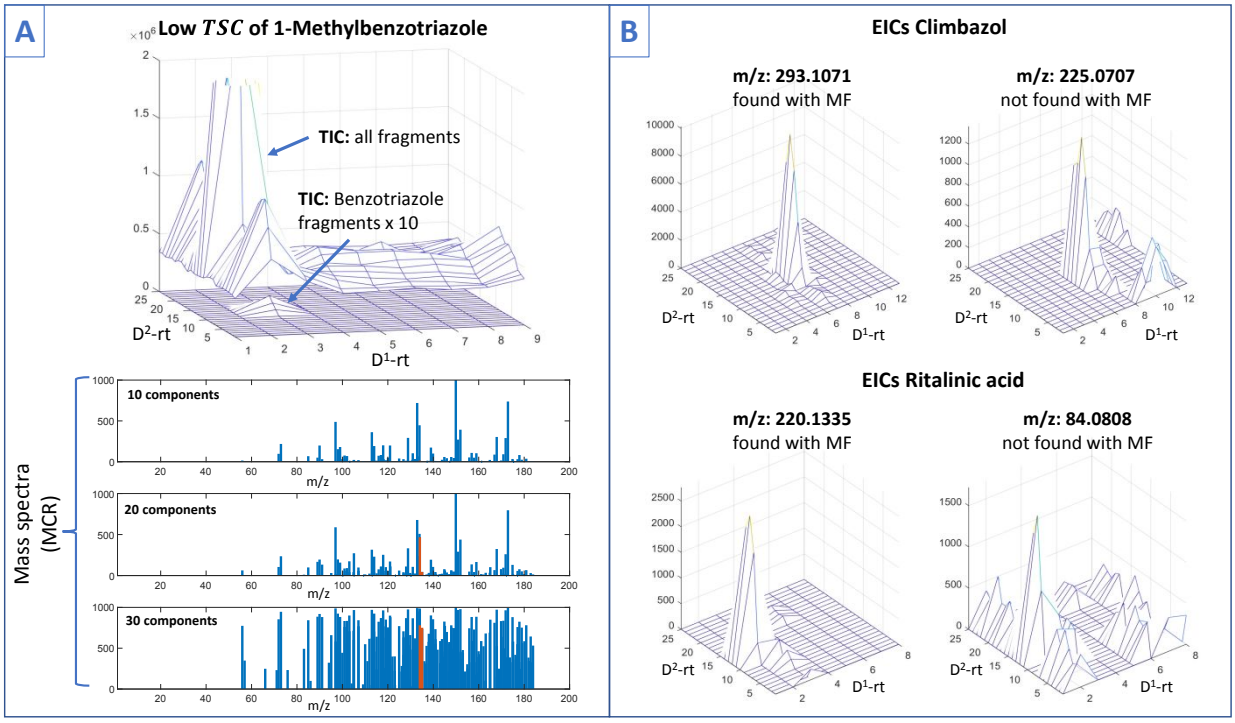


Figure 4*: Visualization of factors causing the observed performance differences between the different data analysis strategies.* ***A****: 1-methylbenzotriazole has very low signal intensity compared with the background. This causes problems for the spectra extraction using MCR.* ***B****: Co-elution and background noise in the EICs of fragments belonging to Climbazol and Ritalinic acid decrease the similarity score calculated in the MF approach.*

In contrast, the MF method was not negatively affected by low *TSC* compounds as the MCR method, since it groups EICs based on their elution profile similarity towards a reference m/z after normalization of each EIC, independently (Table 1). The MF method was challenged if co-elution or background noise in the respective EIC occurred, which was not present in the reference EIC. Figure 4**B** shows the EICs of Climbazol and Ritalinic acid for which the MF method failed to extract satisfactory mass spectra. In the case of Climbazol, co-elution observed in the EIC

decreased the similarity score for the fragment with m/z 225.0707 below the threshold of 0.7 and the identification of Ritalinic acid failed because of too high background noise (m/z 84.081).

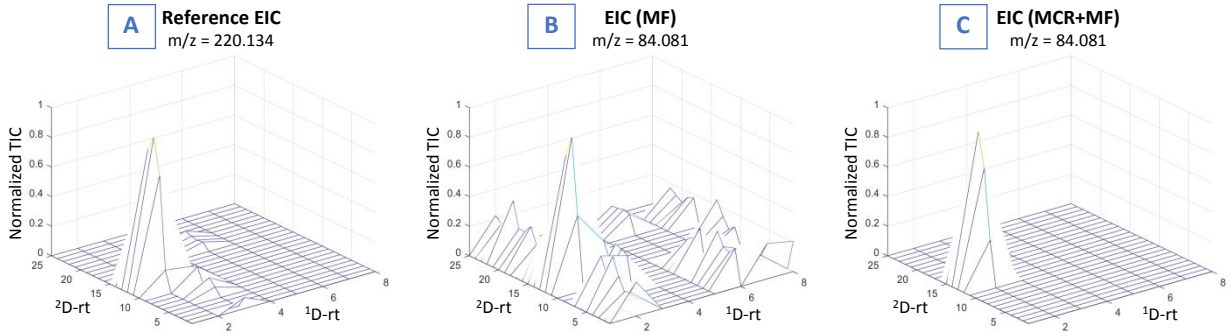


Figure 5**: Visualization of the signal deconvolution achieved applying MF+MCR.* **A: Reference EIC used to calculate similarities with candidate EICs.* **B***: MF alone fails to identify m/z 84.081 as a fragment belonging to Ritalinic acid because of the noisiness of the EIC.* **C***: MCR+MF can separate the pure EIC from the background noise resulting in the identification of Ritalinic acid.*

From the observations made regarding the advantages and limitations of MCR and MF individually, it was hypothesized that using the MF approach to perform data reduction followed by MCR method would be the method able to detect the most compounds. Therefore, the MF method with a lowered threshold was used to improve the TSC prior to performing the MCR, hence being less conservative in including EICs compared to using MF alone. The subsequent MCR step was then applied to resolve co-elution and extract pure mass spectra of the respective suspects which is exemplified in Figure 5. Indeed, the combination of MF and MCR was able to separate co-elution and background noise from the analyte signal resulting in the additional identification of Climbazol and Ritalinic acid as compared to MF without MCR.

Table 1: Results of the suspect screening using the different data analysis strategies.

Compound name	TSC	Detected suspects			
		PAM	MCR	MF	MF+MCR
Fexofenadine	0.0608	1	1	1	1
Lamotrigine	0.0441	1	1	1	1
Losartan carboxylic acid	0.0190	1	1	1	1
Citalopram	0.0165	1	1	1	1
Metoprolol	0.0124	1	1	1	1
2-[4-(Diethylamino)-2-hydroxybenzoyl]benzoic acid	0.0116	1	1	1	1
Galaxolidone	0.0096	0	1	1	1
Amisulpride	0.0073	0	1	1	1
DEET	0.0068	1	1	1	1
2-Ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine	0.0061	0	1	1	1
Ensulizole	0.0061	1	1	1	1
Methadone	0.0048	0	0	1	1
Amitryptilin	0.0047	0	1	1	1
Cetirizine	0.0030	0	1	1	1
Diclofenac	0.0022	0	1	1	1
Losartan carboxaldehyde	0.0014	0	1	1	1
Carbamazepine	0.0013	1	1	1	1
Gabapentin lactam	0.0011	1	0	1	1
Terbutryn	0.0011	0	0	1	1
Clopidrogel carboxylic acid	0.0010	0	0	1	1
Verapamil metabolite	0.0007	0	0	1	1
Venlafaxine	0.0005	0	0	1	1
Climbazol	0.0004	0	0	0	1
1-Methylbenzotriazole	0.0004	0	0	1	1
Ritalinic acid	0.0004	0	0	0	1

3.2 Robustness of the MF and MCR workflows

The quality of an extracted mass spectrum of a suspect was found to be a compromise between removing co-eluting and background ions (false positives) and retaining fragment ions of the suspect (true positives). In the presented workflow, this compromise was optimized with the user defined similarity metric i.e., a higher value resulted in fewer false positives whereas increasing the value too much would decrease the number of true positives. The selectivity of the MF method highly benefits from the combination of the two retention dimensions, which sets it apart from 1D feature based workflows. In Figure 6A, it can be seen that the data reduction achieved is around a

factor 20-30 larger when similarity scores are calculated incorporating both retention dimensions as opposed to calculating similarity scores based only on one of the retention dimensions.

In this study, a similarity score threshold of 0.7 was chosen for extracting mass spectra from the wastewater sample using the MF method. In the following, the effect of choosing higher or lower similarity score thresholds is exemplified using a subset of suspect compounds.

Four suspects are shown in Figure 6B-E representing four different scenarios that we observed with respect to the threshold. The evaluation of all suspects is given in S4. The first scenario shown in Figure 6B represents 8/28 cases in which false positive matches could be filtered out (e.g., Tramadol). False positive matches in this context means that characteristic fragments of a suspect are present in the ${}^{1}t_r \times {}^{2}t_r$ but closer investigation of the EICs showed that the fragments embodied noise and no chemical information. Hence, only looking at the presence of specific fragments in the ${}^{1}t_r \times {}^{2}t_r$ frame without considering the EIC shapes would have led to falsely assuming that a suspect is present in the sample. The selected threshold of 0.7 has a significant safety margin, as none of the false positive matches reaches similarities greater than 0.4. Figure 6C represents 10/28 suspects for which the number of true positives was insensitive to the actual similarity score threshold up to very high threshold values. Even similarity score threshold values of larger than 0.95 would not have reduced the number of true positives significantly, whereas the TSC could be increased to values close to 1 indicating that the mass spectrum has been perfectly purified. On the other hand, Figures 6D-E exemplify the two scenarios that advocate for a conservative similarity score threshold. In Figure 6D it can be seen that for Metropolol the number of true positives declines rapidly when moving to threshold values larger than 0.7. On the other hand, the similarity score threshold was too high to detect a decisive characteristic fragments in the cases of Ritalinic acid and Climbazol which led to not identifying these two suspects with MF at a threshold of 0.7. However, the missed-out EICs yielded similarity scores close to the threshold (0.67 and 0.69 for Ritalinic acid fragment with m/z 84.081 and Climbazol fragment with m/z 225.0707 respectively). In the investigated data set the only observed mode of failure for MF was related to missing out true positives and not due to the noisiness of the extracted mass spectra as it has been the case for PAM and MCR. We conclude that similarity score threshold values in the range from 0.6 to 0.7 should provide good results by filtering out background noise, hence obtaining low background interferent of the extracted mass spectra. If more severe co-elution or higher background noise is expected, it is advisable to use MF with a lower similarity score threshold in combination with a consecutive MCR step.

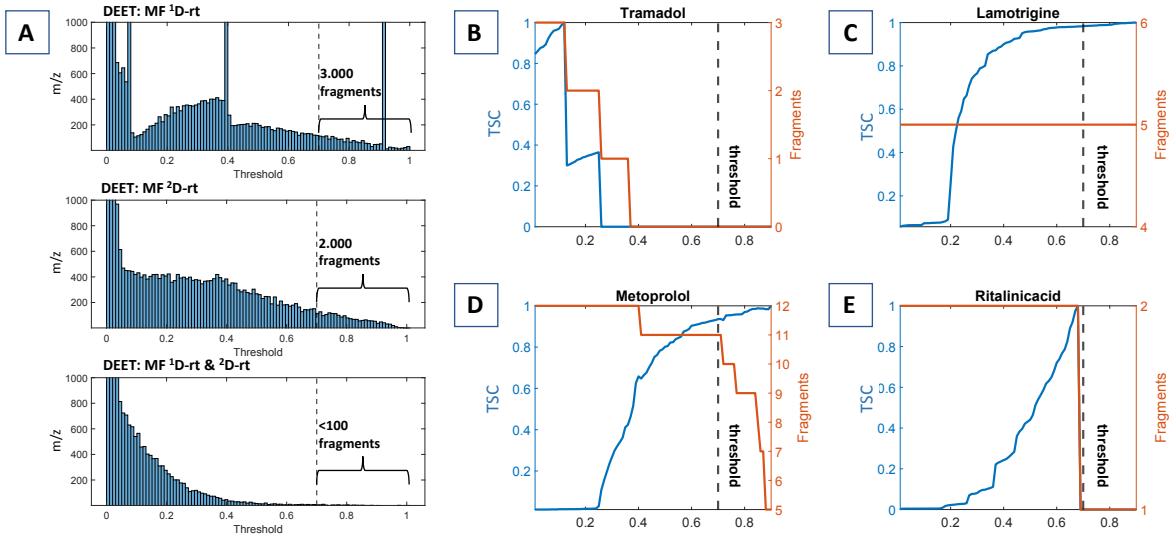


***Figure 6**: Analysis of the MF threshold. **A**: Calculation of similarity metric incorporating information from both retention dimensions provides selectivity gain of a factor 20 to 30 compared to similarity metric calculated on a single retention dimension only. **B-E**: Comparison of the TSC ratio and the number of reference fragments found in MS2 data as a function of the threshold.*

Regarding the robustness of MCR it was observed that inconsistencies between EIC's introduced by the ROI preprocessing can lead to problems beyond the outlined issue of low TSC. We observed elution profile changes for EICs with low intensities specifically at the front or tail of the peak where scan points were missing. This could be caused by improper grouping of mass spectral measurements in the ROI processing, loss of low intensity data points when centroiding the profile data, or due to too low ion counts to be properly measured by the detector. In Figure 7 an example of two EICs that belong to Terbutryn are shown. It was observed that the fragment with a m/z = 158.0485 was only detected in one modulation and that there were missing data points in the front and tail of the peak in the ^{2}D compared to fragment with mz_2 = 186.0810. These artifacts are problematic for MCR (with and without prior MF) because they violate the assumption that the shapes of all EICs belonging to one analyte are the same.[47,48] In case of Terbutryn the consequence was that the EICs having different shapes were modelled by different factors if the models were slightly overfitted.

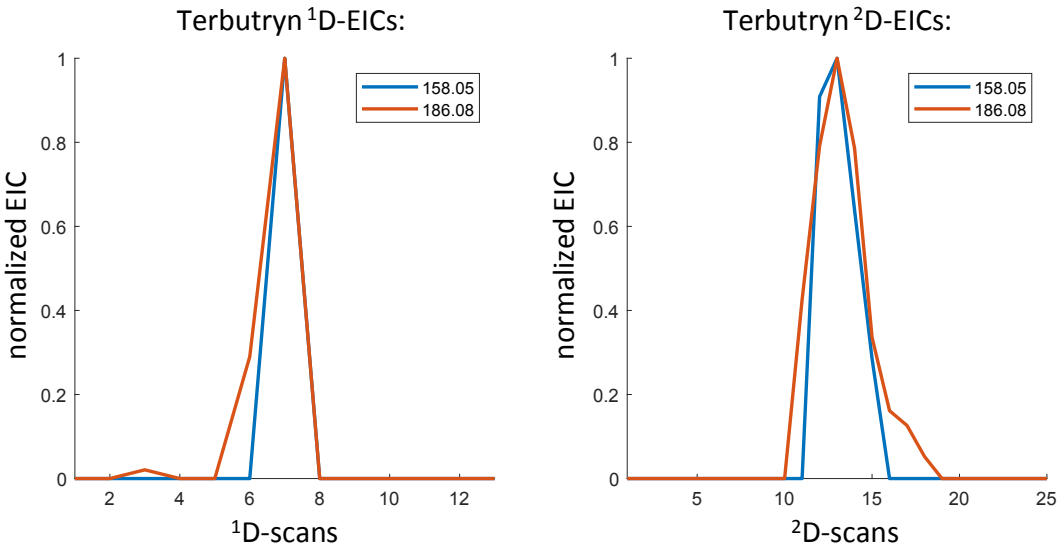


***Figure 7**: Artifacts in the EICs caused by the ROI preprocessing.*

Although, the similarity score calculated in the MF procedure is also negatively affected by the shape differences, it was found to be more robust compared to MCR. In fact, the similarity scores calculated between the two EICs and the reference profile of the $[M+H]^+$ peak were 0.94 and 0.92 (respectively for mz_1 = 158.0485 and mz_2 = 186.0810), which is well above the similarity score threshold of 0.7 used in this study. We observed that smoothing of the data could improve the robustness of the MCR models but on the other hand lowered the specificity of the MF. A detailed description of the effect of smoothing is given in the supporting material S3. Therefore, we propose to apply smoothing, after performing the MF step.

4 Conclusions

In this study we introduced a simple and effective data analysis strategy for the extraction of pure MS2 mass spectra from LC×LC-HRMS measurements for the identification of compounds in suspects screening. The proposed MF method utilizes the elution profiles of the ^{1}D and ^{2}D to filter out EICs that do not belong to the mass spectrum of a compound. While MF alone outperformed MCR for the number of identified suspects, the MF+MCR method emerged as the most effective strategy to overcome the limitations posed by co-elution and background noise as well as low TSC. This approach notably facilitated the confirmation of the presence of 25 suspect compounds in the sample, demonstrating a significant improvement over the other methods. The success of

MF+MCR was attributed to the synergy of MF's robust noise filtration and MCR's ability to resolve co-elution, thus enhancing the quality of the mass spectra obtained.

However, each method's performance was intrinsically linked to factors such as the *TSC* and the extent of co-elution or background noise. Our study suggests that while MCR is sensitive to compounds with low signal intensity, the MF method's performance is primarily affected by the presence of co-elution or background noise in the EICs. The robustness of both MF and MCR methods against these challenges was critically assessed, with a recommendation for a judicious selection of MF threshold values to balance the retention of true positives against the exclusion of false positives.

We found that data smoothing could be a viable solution to increase the robustness of the MCR models in situations where small inconsistencies in EIC peak shapes were observed.

In conclusion, the combined MF+MCR method, with its enhanced signal deconvolution capabilities, stands out as a powerful tool for signal processing of LC×LC-HRMS data. This study provided a strategic framework for selecting and optimizing data analysis workflows in LC×LC-HRMS.

5 Acknowledgements

The authors would like to acknowledge Rasmus Bro for fruitful discussions and revising the manuscript. We would like to acknowledge Giorgio Tomasi for an interesting discussion on data smoothing. Additionally, we would like to acknowledge Jan H. Christensen, Nikoline Juul Nielsen and Selina Tisler for making the wastewater effluent extract available. Furthermore, we would like to thank Jan H. Christensen and Nikoline Juul Nielsen for helping procuring the pulsed elution-LC×LC-HRMS data in collaboration with the last author of this paper.

6 Conflict of Interest

The authors declare no conflict of interest.

7 CRediT statement

Paul-Albert Schneide: Conceptualization, Formal Analysis, Writing – original draft, Writing – review & editing

Oskar Munk Kronik: Conceptualization, Formal Analysis, Writing – original draft, Writing – review & editing

8 References

(1) Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; Thomaidis, N. S.; Bletsou, A.; Zwiener, C.; Ibáñez, M.; Portolés, T.; De Boer, R.; Reid, M. J.; Onghena, M.; Kunkel, U.; Schulz, W.; Guillon, A.; Noyon, N.; Leroy, G.; Bados, P.; Bogialli, S.; Stipaničev, D.; Rostkowski, P.; Hollender, J. Non-Target Screening with High-Resolution Mass Spectrometry: Critical Review Using a Collaborative Trial on Water Analysis. *Anal Bioanal Chem* **2015**, *407* (21), 6237–6255. https://doi.org/10.1007/s00216-015-8681-7.

(2) Brüggen, S.; Schmitz, O. J. A New Concept for Regulatory Water Monitoring Via High-Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry. *J Anal Test* **2018**, *2* (4), 342–351. https://doi.org/10.1007/s41664-018-0081-5.

(3) Tisler, S.; Engler, N.; Jørgensen Blichert, M.; Kilpinin, K.; Tomasi, G.; Christensen, J. H. From Data to Reliable Conclusions: Identification and Comparison of Persistent Micropollutants and Transformation Products in 37 Wastewater Samples by Non-Target Screening Prioritization. *Water Res* **2022**, *219*. https://doi.org/doi: 10.1016/j.watres.2022.118599.

(4) Schollée, J. E.; Bourgin, M.; von Gunten, U.; McArdell, C. S.; Hollender, J. Non-Target Screening to Trace Ozonation Transformation Products in a Wastewater Treatment Train Including Different Post-Treatments. *Water Res* **2018**, *142*, 267–278. https://doi.org/10.1016/j.watres.2018.05.045.

(5) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; Sarvepalli, A.; Zhang, Z.; Fleischauer, M.; Dührkop, K.; Wesner, M.; Hoogstra, S. J.; Rudt, E.; Mokshyna, O.; Brungs, C.; Ponomarov, K.; Mutabdžija, L.; Damiani, T.; Pudney, C. J.; Earll, M.; Helmer, P. O.; Fallon, T. R.; Schulze, T.; Rivas-Ubach, A.; Bilbao, A.; Richter, H.; Nothias, L. F.; Wang, M.; Orešič, M.; Weng, J. K.; Böcker, S.; Jeibmann, A.; Hayen, H.; Karst, U.; Dorrestein, P. C.; Petras, D.; Du, X.; Pluskal, T. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nature Biotechnology*. Nature Research April 1, 2023, pp 447–449. https://doi.org/10.1038/s41587-023-01690-2.

(6) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. XCMS2: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. *Anal Chem* **2008**, *80* (16), 6382–6389. https://doi.org/10.1021/ac800795f.

(7) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; Vandergheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis. *Nat Methods* **2015**, *12* (6), 523–526. https://doi.org/10.1038/nmeth.3393.

(8) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. PatRoon: Open Source Software Platform for Environmental Mass Spectrometry Based Non-Target Screening. *J Cheminform* **2021**, *13* (1). https://doi.org/10.1186/s13321-020-00477-w.

(9) Bittremieux, W.; Wang, M.; Dorrestein, P. C. The Critical Role That Spectral Libraries Play in Capturing the Metabolomics Community Knowledge. *Metabolomics*. Springer December 1, 2022. https://doi.org/10.1007/s11306-022-01947-y.

(10) Wang, M.; Jarmusch, A. K.; Vargas, F.; Aksenov, A. A.; Gauglitz, J. M.; Weldon, K.; Petras, D.; da Silva, R.; Quinn, R.; Melnik, A. V.; van der Hooft, J. J. J.; Caraballo-Rodríguez, A. M.; Nothias, L. F.; Aceves, C. M.; Panitchpakdi, M.; Brown, E.; Di Ottavio, F.; Sikora, N.; Elijah, E. O.; Labarta-Bajo, L.; Gentry, E. C.; Shalapour, S.; Kyle, K. E.; Puckett, S. P.; Watrous, J. D.; Carpenter, C. S.; Bouslimani, A.; Ernst, M.; Swafford, A. D.; Zúñiga, E. I.; Balunas, M. J.;

Klassen, J. L.; Loomba, R.; Knight, R.; Bandeira, N.; Dorrestein, P. C. Mass Spectrometry Searches Using MASST. *Nat Biotechnol* **2020**, *38* (1), 23–26. https://doi.org/10.1038/s41587-019-0375-9.

(11) Renner, G.; Reuschenbach, M. Critical Review on Data Processing Algorithms in Non-Target Screening: Challenges and Opportunities to Improve Result Comparability. *Analytical and Bioanalytical Chemistry*. Springer Science and Business Media Deutschland GmbH July 1, 2023, pp 4111–4123. https://doi.org/10.1007/s00216-023-04776-7.

(12) Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly Sensitive Feature Detection for High Resolution LC/MS. *BMC Bioinformatics* **2008**, *9*. https://doi.org/10.1186/1471-2105-9-504.

(13) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal Chem* **2017**, *89* (17), 8689–8695. https://doi.org/10.1021/acs.analchem.7b01069.

(14) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal Chem* **2012**, *84* (1), 283–289. https://doi.org/10.1021/ac202450g.

(15) Heuckeroth, S.; Damiani, T.; Smirnov, A.; Mokshyna, O.; Brungs, C.; Korf, A.; Smith, J. D.; Stincone, P.; Dreolin, N.; Nothias, L. F.; Hyötyläinen, T.; Orešič, M.; Karst, U.; Dorrestein, P. C.; Petras, D.; Du, X.; van der Hooft, J. J. J.; Schmid, R.; Pluskal, T. Reproducible Mass Spectrometry Data Processing and Compound Annotation in MZmine 3. *Nat Protoc* **2024**. https://doi.org/10.1038/s41596-024-00996-y.

(16) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data Analysis Strategies for Targeted and Untargeted LC-MS Metabolomic Studies: Overview and Workflow. *TrAC - Trends in Analytical Chemistry* **2016**, *82*, 425–442. https://doi.org/10.1016/j.trac.2016.07.004.

(17) Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Anal Chem* **2020**, *92* (2), 1898–1907. https://doi.org/10.1021/acs.analchem.9b04095.

(18) Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z. J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M. A Lipidome Atlas in MS-DIAL 4. *Nat Biotechnol* **2020**, *38* (10), 1159–1163. https://doi.org/10.1038/s41587-020-0531-2.

(19) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. XCMS2: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. *Anal Chem* **2008**, *80* (16), 6382–6389. https://doi.org/10.1021/ac800795f.

(20) Pérez-López, C.; Oró-Nolla, B.; Lacorte, S.; Tauler, R. Regions of Interest Multivariate Curve Resolution Liquid Chromatography with Data-Independent Acquisition Tandem Mass Spectrometry. *Anal Chem* **2023**, *95* (19), 7519–7527. https://doi.org/10.1021/acs.analchem.2c05704.

(21) Navarro-Reig, M.; Jaumot, J.; Tauler, R. An Untargeted Lipidomic Strategy Combining Comprehensive Two-Dimensional Liquid Chromatography and Chemometric Analysis. *J Chromatogr A* **2018**, *1568*, 80–90. https://doi.org/10.1016/j.chroma.2018.07.017.

(22) Amigo, J. M.; Skov, T.; Bro, R.; Coello, J.; Maspoch, S. Solving GC-MS Problems with PARAFAC2. *TrAC - Trends in Analytical Chemistry* **2008**, *27* (8), 714–725. https://doi.org/10.1016/j.trac.2008.05.011.

(23) Kronik, O. M.; Liang, X.; Nielsen, N. J.; Christensen, J. H.; Tomasi, G. Obtaining Clean and Informative Mass Spectra from Complex Chromatographic and High-Resolution All-Ions-Fragmentation Data by Nonnegative Parallel Factor Analysis 2. *J Chromatogr A* **2022**, *1682*, 463501. https://doi.org/10.1016/j.chroma.2022.463501.

(24) Gorrochategui, E.; Jaumot, J.; Tauler, R. ROIMCR: A Powerful Analysis Strategy for LC-MS Metabolomic Datasets. *BMC Bioinformatics* **2019**, *20* (1), 256. https://doi.org/10.1186/s12859-019-2848-8.

(25) Schneide, P.; Bro, R.; Gallagher, N. B. Shift-invariant Tri-linearity—A New Model for Resolving Untargeted Gas Chromatography Coupled Mass Spectrometry Data. *J Chemom* **2023**, *37* (8). https://doi.org/10.1002/cem.3501.

(26) Hoggard, J. C.; Siegler, W. C.; Synovec, R. E. Toward Automated Peak Resolution in Complete GC × GC-TOFMS Chromatograms by PARAFAC. *J Chemom* **2009**, *23* (7–8), 421–431. https://doi.org/10.1002/cem.1239.

(27) Hoggard, J. C.; Synovec, R. E. Parallel Factor Analysis (PARAFAC) of Target Analytes in GC × GC-TOFMS Data: Automated Selection of a Model with an Appropriate Number of Factors. *Anal Chem* **2007**, *79* (4), 1611–1619. https://doi.org/10.1021/ac061710b.

(28) Hoggard, J. C.; Synovec, R. E. Automated Resolution of Nontarget Analyte Signals in GC x GC-TOFMS Data Using Parallel Factor Analysis. *Anal Chem* **2008**, *80* (17), 6677–6688. https://doi.org/10.1021/ac800624e.

(29) Jonsson, P.; Johansson, A. I.; Gullberg, J.; Trygg, J.; A, J.; Grung, B.; Marklund, S.; Sjöström, M.; Antti, H.; Moritz, T. High-Throughput Data Analysis for Detecting and Identifying Differences between Samples in GC/MS-Based Metabolomic Analyses. *Anal Chem* **2005**, *77* (17), 5635–5642. https://doi.org/10.1021/ac050601e.

(30) Tomasi, G.; Acar, E.; Bro, R. *Multilinear Models, Iterative Methods*, Second Edi.; Elsevier, 2020; Vol. 2. https://doi.org/10.1016/b978-0-12-409547-2.14609-8.

(31) Abdollahi, H.; Tauler, R. Uniqueness and Rotation Ambiguities in Multivariate Curve Resolution Methods. *Chemometrics and Intelligent Laboratory Systems* **2011**, *108* (2), 100–111. https://doi.org/10.1016/j.chemolab.2011.05.009.

(32) Tauler, R. Multivariate Curve Resolution Applied to Second Order Data. **1995**, *7439* (95).

(33) Harshman, R. A. *PARAFAC2: Mathematical and Technical Notes.*; 1972; Vol. 22.

(34) de Juan, A.; Tauler, R. Multivariate Curve Resolution: 50 Years Addressing the Mixture Analysis Problem – A Review. *Anal Chim Acta* **2021**, *1145*, 59–78. https://doi.org/10.1016/j.aca.2020.10.051.

(35) Stoll, D. R.; Lhotka, H. R.; Harmes, D. C.; Madigan, B.; Hsiao, J. J.; Staples, G. O. High Resolution Two-Dimensional Liquid Chromatography Coupled with Mass Spectrometry for Robust and Sensitive Characterization of Therapeutic Antibodies at the Peptide Level. *J Chromatogr B Analyt Technol Biomed Life Sci* **2019**, *1134–1135* (October), 121832. https://doi.org/10.1016/j.jchromb.2019.121832.

(36) Venter, P.; Muller, M.; Vestner, J.; Stander, M. A.; Tredoux, A. G. J.; Pasch, H.; De Villiers, A. Comprehensive Three-Dimensional LC × LC × Ion Mobility Spectrometry Separation Combined with High-Resolution MS for the Analysis of Complex Samples. *Anal Chem* **2018**, *90* (19), 11643–11650. https://doi.org/10.1021/acs.analchem.8b03234.

(37) Saint Germain, F. M.; Faure, K.; Saunier, E.; Lerestif, J. M.; Heinisch, S. On-Line 2D-RPLC x RPLC – HRMS to Assess Wastewater Treatment in a Pharmaceutical Plant. *J Pharm Biomed Anal* **2022**, *208*, 114465. https://doi.org/10.1016/j.jpba.2021.114465.

(38) Muller, M.; Tredoux, A. G. J.; de Villiers, A. Predictive Kinetic Optimisation of Hydrophilic Interaction Chromatography × Reversed Phase Liquid Chromatography Separations: Experimental Verification and Application to Phenolic Analysis. *J Chromatogr A* **2018**, *1571*, 107–120. https://doi.org/10.1016/j.chroma.2018.08.004.

(39) Pirok, B. W. J.; Pous-Torres, S.; Ortiz-Bolsico, C.; Vivó-Truyols, G.; Schoenmakers, P. J. Program for the Interpretive Optimization of Two-Dimensional Resolution. *J Chromatogr A* **2016**, *1450*, 29–37. https://doi.org/10.1016/j.chroma.2016.04.061.

(40) Hugelier, S.; Devos, O.; Ruckebusch, C. Constraining Shape Smoothness in Multivariate Curve Resolution–Alternating Least Squares. *J Chemom* **2015**, *29* (8), 448–456. https://doi.org/10.1002/cem.2724.

(41) Cain, C. N.; Trinklein, T. J.; Ochoa, G. S.; Synovec, R. E. Tile-Based Pairwise Analysis of GC × GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification. *Anal Chem* **2022**, *94* (14), 5658–5666. https://doi.org/10.1021/acs.analchem.2c00223.

(42) Ochoa, G. S.; Sudol, P. E.; Trinklein, T. J.; Synovec, R. E. Class Comparison Enabled Mass Spectrum Purification for Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry. *Talanta* **2022**, *236*. https://doi.org/10.1016/j.talanta.2021.122844.

(43) Kronik, O. M.; Christensen, J. H.; Nielsen, N. J. Instrumental and Theoretical Advancements in Pulsed Elution-LC × LC: Investigation of Pulse Parameters and Application to Wastewater Effluent. *J Chromatogr A* **2024**, *1730*, 465079. https://doi.org/10.1016/j.chroma.2024.465079.

(44) Tisler, S.; Pattison, D. I.; Christensen, J. H. Correction of Matrix Effects for Reliable Non-Target Screening LC–ESI–MS Analysis of Wastewater. *Anal Chem* **2021**, *93* (24), 8432–8441. https://doi.org/10.1021/acs.analchem.1c00357.

(45) Bader, B. W.; Kolda, T. G. Efficient MATLAB Computations with Sparse and Factored Tensors. *SIAM Journal on Scientific Computing* **2008**, *30* (1), 205–231. https://doi.org/10.1137/060676489.

(46) Kristupas Tikuisis. Peaks2. MATLAB Central File Exchange April 12, 2023.

(47) Cattell, R. B. “Parallel Propprtional Profiles” and Other Principles for Determining the Choice of Factors by Rotation. *Psychometrika* **1944**, *9* (4), 267–283.

(48) Lawton, W. H.; Sylvestre, E. A. Self Modeling Curve Resolution. *Technometrics* **1971**, *13* (3), 617. https://doi.org/10.2307/1267173.

S1: Mathematical description of chemometric methods

MCR:

MCR is a chemometric technique used to decompose complex data sets into pure component spectra and their associated concentration profiles. Mathematically, MCR models a data matrix **D** as the product of two matrices, a concentration profile matrix **C** and a pure component spectra matrix **S**, plus a residual matrix **E** that accounts for the model's unexplained variance:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E}$$

Where:

- **D** is the data matrix of concatenated elution profiles with dimensions $kl \times n$, where k is the number of modulations in ^{1}D, l is the number of scans in ^{2}D, and n is the number of mass channels (e.g., m/z values).

- **C** is an $kl \times p$ matrix, where p is the number of latent variables. Each column represents the concatenated elution profiles of a component across the samples.

- **S** is an $n \times p$ matrix, with each column representing the pure mass spectrum of a component.

- **E** is the residual matrix of the same dimensions as **D**, representing noise and information not captured by the model.

The MCR problem is often ill-posed and requires constraints for a unique solution, such as non-negativity of the concentration and spectra matrices, unimodality of the concentration profiles, or closure constraints. The iterative optimization of MCR typically involves alternating least squares (ALS) to estimate **C** and **S** until convergence is achieved. For MCR, the selection of the number of components is critical. This is often done using methods like cross-validation, scree plots, or core consistency diagnostics. In practice, these models are applied to small regions of the chromatographic data set (windowing) to manage computational complexity and model accuracy.

S2: Detailed description of the pre-processing

ROI

An inhouse developed ROI algorithm was used for the binning of the m/z direction. The ROI algorithm takes the centroided raw data as input and performs a special form of binning in the mass direction to return aligned mass traces as a list of potential features and discard uninformative mass channels (compare **Figure 1**). It takes three input arguments: the mass tolerance, the minimum intensity, and the minimum peak width (number of consecutive scans that form a peak). The mass tolerance is usually set as the instruments mass accuracy and defines the width of the bins in mass direction. The minimum intensity defines a lower threshold to filter out mass values with intensity < than the threshold.The minimum intensity should be set lower than the baseline signal for the lowest intensity analyte of interest, however, often it is set as a compromise between the latter and to decrease processing time and memory footprint on the computer. Finally, the peak

width describes the minimum number of consecutive points in a bin for it to be considered as a feature (Figure 1 step 2). In each bin, the mass values are aligned to a single mass value, which can be a global average or a weighted average (e.g., weighted by the relative intensity of each mass value in a respective bin). In our study, the noise threshold was set to 0 to avoid excluding low abundant peak fragments. The mass error has been varied between 10 ppm and 16 ppm to test the sensitivity of the results towards this parameter. The minimum peak width has been set to 4 scans i.e. 2.5 seconds.

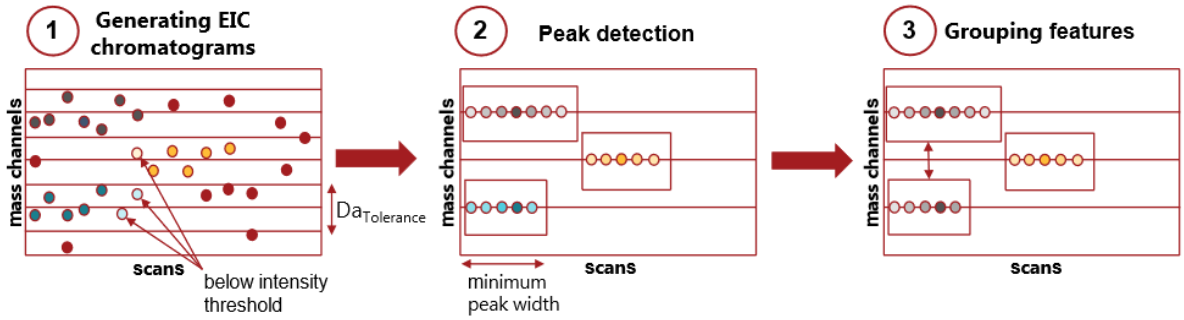


Figure 1: ROI binning exemplified. 1. Based on the centroided raw data mass scans within a specified mass tolerance are aligned to a specific mass. Mass scans with intensity below the user defined noise threshold are discarded. It is shown that discarding low intensity scans can lead to the distortion of EICs because mass scans belonging to an analyte (light green and light yellow) may be discarded 2. The aligned masses are defined as features if there are at least as many consecutive non-zero scans as defined by the minimum peak width parameter. 3. Software packages like mzMine and XCMS perform a similarity-based grouping of the features to reconstruct component mass spectra. This feature is only available for 1D-chromatography.

S2: Data analysis strategies (DAS)

DAS1 Peak Apex method (PAM)

The peak apex method is the most common procedure for extracting mass spectra of the suspect. It includes two steps: 1) finding the coordinates of the peak apex in the EIC of an $[M+H]^+$ ion associated with a suspect compound in MS1 and 2) extracting the mass spectra from MS1 and MS2 at these specific coordinates. These two steps are exemplified for the suspect Fexofenadine in **Figure 2**A-C. In **Figure 2**B only the reduced size TIC of MS1 is shown for better visualization, however, **Figure 2**D shows the mass spectra extracted from MS1 and MS2 using PAM. In the case of Fexofenadine, the extracted mass spectrawere not severely affected by baseline or co-eluting compounds because of the high signal-to-noise ratio (S/N) of this compound. The PAM procedure

is challenged if the S/N is low or if co-elution is occurring. In these cases PAM will not yield clean mass spectra as outlined in section 3.1 of the paper.

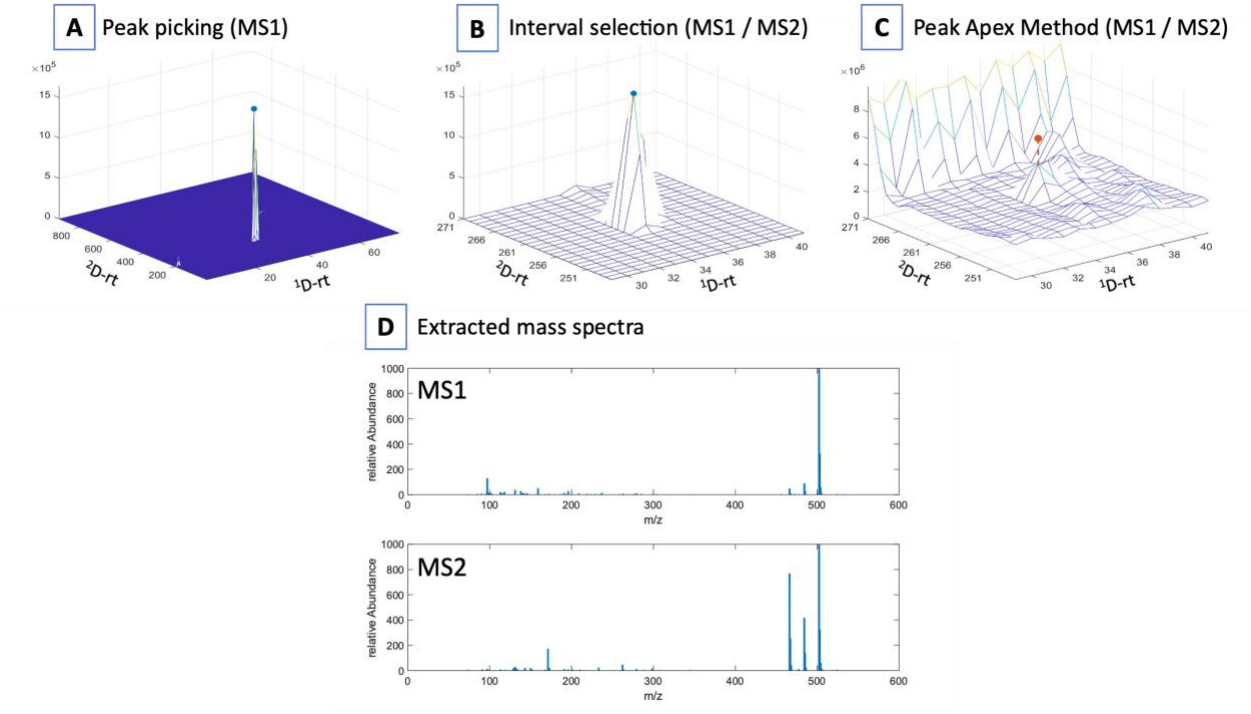


Figure 2: Description of the peak apex method exemplified. A: Peak picking in the 2D EIC of the [M+H]+ ion of the suspect Fexofenadine to find the coordinates of the peak apex.. B: Visualization of the selected ${}^{1}t_r \times {}^{2}t_r$ frame surrounding the apex of the $[M+H]^+$ion. C: Component mass spectra are extracted at the peak apex position from the MS1 / MS2 TICs. Visualized is the MS1 TIC. D: Extracted component mass spectra (MS1 and MS2) for Fexofenadine.

DAS2 MCR

MCR was applied only to the reduced size three-way arrays of size $(k \times l \times n)$, selected around the apex of the $[M+H]^+$ ion in MS1 to reduce the complexity of the data deconvoluted. The dimensions k and l are the number of datapoints selected in ^{1}D and ^{2}D, respectively, which have been fixed to 13 (peak apex $\pm$ 6 datapoints) and 25 datapoints (peak apex $\pm$ 12 datapoints), respectively. The dimension n denotes the number of mass fragments that have non-zero intensity in the ${}^{1}t_r \times {}^{2}t_r$ frame of size $k \times l$. Thus, n varies from suspect to suspect and differs as well between MS1 and MS2 for the ${}^{1}t_r \times {}^{2}t_r$ frame. The MCR models were fitted to the unfolded three-way array **D** with size $(kl \times n)$, as described in **S1**. For the ease of implementation, it was decided to fit MCR

models on the MS1 and MS2 data sets individually and not concatenating MS1 and MS2 via data fusion. It was argued by Tauler et al., that concatenating MS1 and MS2 will reduce the noisiness of the MS2 spectra, because MS1 has better SNR. Our evaluation of the TSC, which can be seen as an alternative formulation for the SNR, (compare *2 Materials and methods*, Eq. (1)) of the reference fragments in MS1 and MS2 did not support this hypothesis. As outlined in section 3.1 in the paper, we found that the performance of the MCR method was good for high TSC values but worsened for low TSC values. In **Figure 3** is the TSC of the reference fragments in MS1 and MS2 for all found suspects visualized. For better visualization, the y-axis has a logarithmic scale (the smaller the TSC, the more negative are the values). **Figure 3** clearly shows that although in some cases MS1 has a higher TSC (this is the case, when the blue bar is less negative than the orange bar), in most cases the TSC of the MS1 is lower than the TSC of MS2. This means that fusing MS1 and MS2 is not guaranteed to reduce the noisiness of MS2 spectra but could also have the reverse effect, since the TSC shows that the contribution of a suspect to the sum-of-squares is equivalent in the MS1 and MS2. Therefore, we do not expect, that our evaluation of the performance of the MCR method was influenced by the decision to calculate the MCR models on MS1 and MS2 individually.

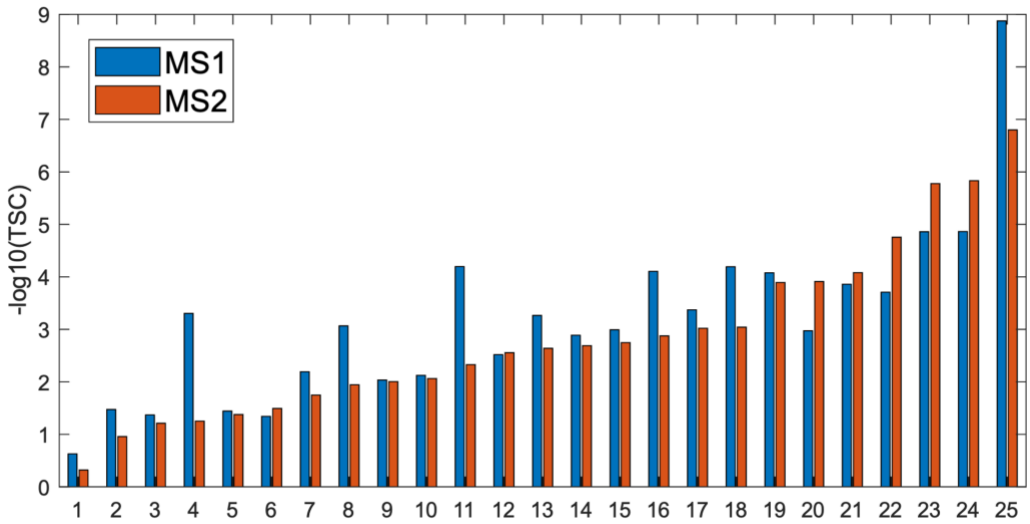


Figure 3: Comparison of the total signal contribution (TSC) for MS1 and MS2. The TSC values have been scaled by taking the negative decadic logarithm, hence high values indicate low TSC values. In many cases the TSC of MS1 is lower than the TSC of MS2. In these cases, the noisiness of the MS2 mass spectra would not be reduced by doing the MCR on the fused MS1 / MS2 data.

The required complexity of the model (number of components) was selected by rank analysis of the unfolded matrices $(kl \times n)$, $(kn \times l)$, and $(ln \times k)$. The number of principle components varied largely across suspects and, for a given suspect as well, across the differently unfolded matrices. For a given suspect, we decided to take the smallest number of components that was found by the rank analysis of the differently unfolded matrices. However, we manually inspected the models and adjusted the number of components, if the model was overfitted or underfitted.

Each MCR model was initialized with random values and fitted five times until convergence. The model with the best lack-of-fit statistic was selected (best out of five). The convergence criterion was set to 1e-08 and the maximum number of iterations was set to 1000.

From the resulting MCR model for a given suspect, the component best describing the suspect was selected by comparing the intensity values of the $[M+H]^+$ ion in the modeled mass spectra **S** and selecting the component which had the highest intensity for the $[M+H]^+$ ion (this does not imply that the $[M+H]^+$ ion needs to have the highest abundance of all m/z in the selected mass spectrum) . This procedure is for example shown in **Figure 4**B, where the results of two MCR models (MS1 and MS2) are shown. The models were fitted to the MS1 and MS2 data shown in **Figure 4**A, which is the ${}^1t_r \times {}^2t_r$ frame surrounding the $[M+H]^+$ ion of Fexofenadine. The MS2 spectrum of the second component was selected for comparison against the reference spectra, because it has the highest intensity of the $[M+H]^+$ ion mass. The MCR results of the decomposition of the MS1 data were used to confirm that the selected component indeed models the analyte signal, based on the 2D-TIC of the respective component.

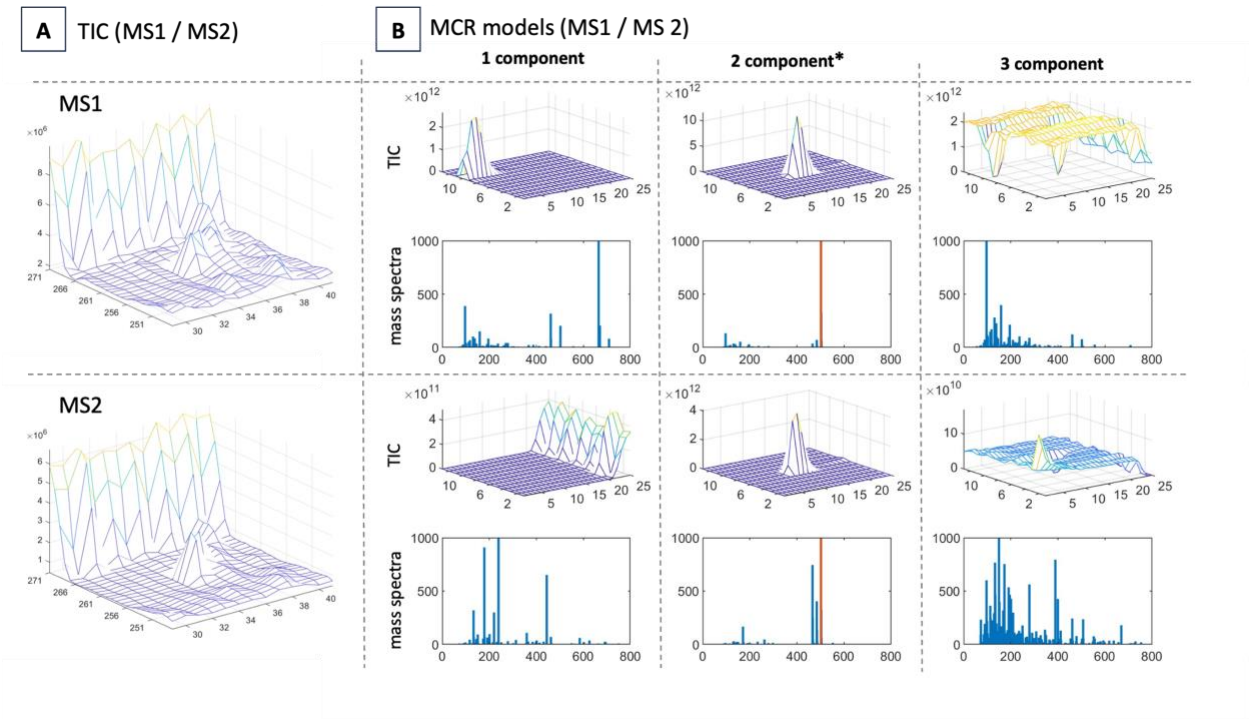


Figure 4: Results of MCR models calculated on the MS1 and MS2 data for Fexofenadine. A: Visualization of the MS1 and MS2 TIC data. B: Results of three component models fitted to the MS1 and MS2 data. Components 1 and 3 are modelling co-eluting peaks and background. Component 2 is modelling the elution profile and mass spectra of Fexofenadine. The MS2 mass spectrum was used for characterization via comparison with literature spectra.

.

DAS3 Mass Filtering (MF)

In this study, we propose a simple workflow to efficiently extract clean analyte mass spectra based on a similarity-based filtering of mass fragments. The first step of this procedure involves the calculation of the reference elution profiles ${}_{ref}^{1}D$ and ${}_{ref}^{2}D$ according to the equations (1) and (2):

(1) $${}_{ref}^{1}D = \frac{\sum_{i=l_1:l_{25}} {}_{i}^{2}D}{\sqrt{\sum_{j=k_1:k_{13}} \left(\sum_{i=l_1:l_{25}} {}_{i}^{2}D\right)^2}}$$

(2) $${}_{ref}^{2}D = \frac{\sum_{j=k_1:k_{13}} {}_{j}^{1}D}{\sqrt{\sum_{i=l_1:l_{25}} \left(\sum_{j=k_1:k_{13}} {}_{j}^{1}D\right)^2}}$$

Where:

- ${}_{ref}^{1}D$ is a normalized row vector of the size $1 \times k$, with k being the number of data points of the retention time interval surrounding the suspect peak in ^{1}D (fixed to be 13 in our study)
- ${}_{ref}^{2}D$ is a normalized column vector of the size $l \times 1$, with l being the number of data points of the retention time interval surrounding the suspect peak in ^{2}D (fixed to be 25 in our study)
- ${}_{j}^{1}D$ is the jth column vector of the data matrix D of size $l \times k$.
- ${}_{i}^{2}D$ is the ith column vector of the data matrix D of size $l \times k$.
- D contains the intensity values of the two-dimensional EIC of the $[M+H]^+$ ion for a given suspect in MS1

After the calculation of the reference elution profiles, candidate elution profiles ${}_{candidate}^{1}D$ and ${}_{candidate}^{2}D$ are calculated for all mz-traces in the selected $l\,x\,k$ window in MS1 and MS2, following the same procedure. Subsequently, the cosine angles ${}_{candidate}^{1}S$ and ${}_{candidate}^{2}S$ between each candidate elution profile and the reference elution profile are calculated by taking the scalar product between both vectors:

(3) $${}_{candidate}^{1}S = {}_{ref}^{1}D \, {}_{candidate}^{1}D^{T}$$

(4) $${}_{candidate}^{2}S = {}_{ref}^{2}D^{T} \, {}_{candidate}^{2}D$$

Where:

- ${}_{candidate}{}^{1}D$ is a matrix of size $n \times k$ holding the candidate elution profiles calculated analogous to Eq. (1) for n mz-tracees in its rows
- ${}_{candidate}{}^{2}D$ is a matrix of size $l \times n$ holding the candidate elution profiles calculated analogous to Eq. (2) for n mass fragments in its columns
- ${}_{candidate}{}^{1}S$ and ${}_{candidate}{}^{2}S$ are column vectors of size $n \times 1$ holding the cosine angles calculated between the reference elution profiles and the candidate elution profiles

The total mass filtering score ${}_{total}{}^{12}S$ is calculated by taking the Hadamard product (entry-wise multiplication) between ${}_{candidate}{}^{1}S$ and ${}_{candidate}{}^{2}S$:

(5) $$ {}_{total}{}^{12}S = {}_{candidate}{}^{1}S \circ {}_{candidate}{}^{2}S $$

The final mass filtering step is simply achieved by discarding all fragments which have a total similarity score below a defined threshold Tr (e.g., 0.7).

A visual example of the mass filtering procedure is given in **Figure 5**, showing the construction of reference profiles from the EIC-MS1 (**Figure 5**A), and the results of the similarity calculation for two selected MS2 fragments (**Figure 5**B). While the blue and orange fragments with mz 502.298 and 484.29 belong to Fexofenadine, the yellow fragment with mz 171.118 stems from background noise. The total similarity score for the two fragments belonging to Fexofenadine are well above the threshold of 0.7 while the total similarity score of the background noise is significantly lower than the threshold.

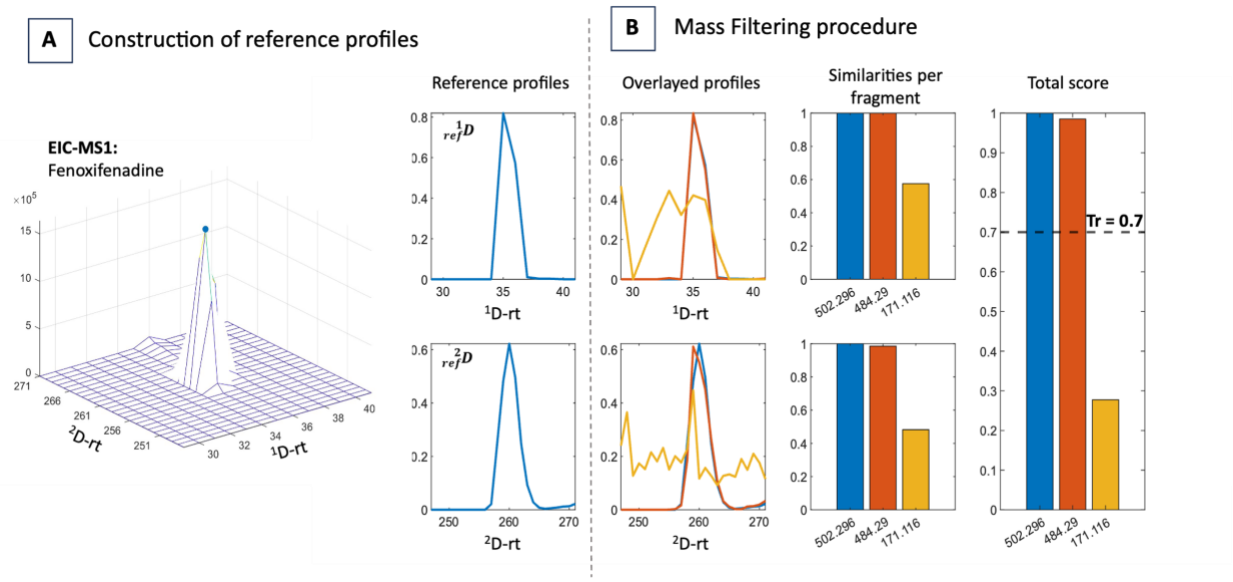
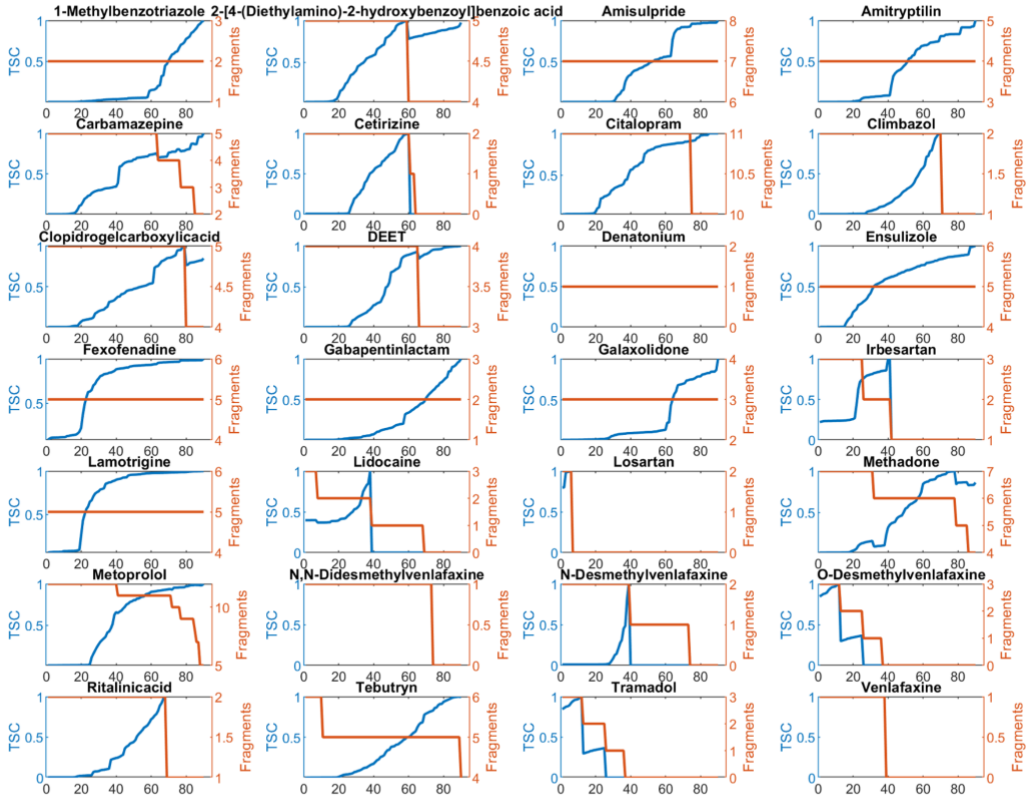


Figure 5: Mass filtering procedure exemplified for Fexofenadine. ***A****: Reference profiles are created based on the EIC of the [M+H]⁺ ion in MS1.* ***B****: The similarity of two selected candidate mass fragments are calculated. Overlays of the candidate mass fragments visually show the similarity of the blue and orange fragment with the reference fragment (left). The similarities are calculated for each retention dimension individually (middle). To enhance the discriminative performance of the procedure, the total score is calculated (right). The blue and orange fragments have a total score above the threshold and are kept for the reconstruction of the component mass spectrum, while the yellow fragment is discarded.*

S3: Raw data table suspect detection

Suspect name	Identification	Found	Similarity				Shared fragments				Library hits			
			PAM	MCR	MF	MF+MC	PAM	MCR	MF	MF+MC	PAM	MCR	MF	MF+MC
\1Methylbenzotriazole	Literature	1	0.30	0.01	1.00	1.00	2	2	2	2	2	2	2	2
\2Ethylidene15dimethyl33diphenylpyrrolidineTPfromMethadione	MASST	1	0.47	0.86	0.70	0.89	3	8	8	8	3	3	3	3
\24Diethylamino2hydroxybenzoylbenzoicacid	Literature	1	0.78	0.80	0.81	0.83	5	5	4	4	1	1	1	1
\Amisulpride	Literature	1	0.47	0.74	0.78	0.77	7	7	7	7	1	1	1	1
\Amitryptilin	Literature	0	0.48	0.58	0.68	0.86	4	4	4	4	1	1	1	1
\Amitryptilin	MASST	1	0.00	0.45	0.64	0.55	0	3	5	3	0	2	11	6
\Carbamazepine	Literature	1	0.67	0.81	0.82	0.65	5	5	4	4	1	1	1	1
\Cetirizine	Literature	0	0.01	0.76	0.00	0.98	2	2	0	2	1	1	1	1
\Cetirizine	Literature	1	0.00	0.71	0.81	0.92	0	4	3	6	0	10	8	10
\Citalopram	Literature	1	0.71	0.88	0.72	0.98	4	4	4	4	1	1	1	1
\Climbazol	Literature	1	0.15	0.01	0.97	0.97	2	2	1	2	1	1	1	1
\Climbazol	MASST	0	0.00	0.00	0.00	0.48	0	0	0	3	0	0	0	3
\Clopidrogelcarboxylicacid	Literature	1	0.20	0.01	0.55	0.49	5	5	5	5	1	1	1	1
\Clopidrogelcarboxylicacid	MASST	1	0.00	0.00	0.92	0.90	0	0	7	7	1	1	1	1
\DEET	Literature	1	0.72	0.95	0.97	0.97	3	3	2	3	4	4	4	3
\Denatonium	Literature	0	0.12	0.11	0.42	1.00	1	1	1	1	1	1	1	1
\Denatonium	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Diclofenac	Literature	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Diclofenac	MASST	1	0.00	0.81	0.86	0.86	0	3	3	3	0	6	6	6
\Ensulizole	Literature	1	0.83	0.83	0.90	0.88	5	5	5	5	1	1	1	1
\Fexofenadine	Literature	1	0.62	0.64	0.63	0.65	5	5	5	5	1	1	1	1
\FexofenadineMinus2H	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\FexofenadineMinusCH2	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\FexofenadinePlusCH2	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\FexofenadinePlusO	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Gabapentinlactam	Literature	1	0.51	0.31	1.00	1.00	2	2	2	2	1	1	1	1
\Galaxolidone	Literature	1	0.35	0.52	0.61	0.64	3	3	3	3	1	1	1	1
\Irbesartan	Literature	0	0.18	0.16	0.15	0.16	3	3	1	1	1	1	1	1
\Irbesartan	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Lamotrigine	Literature	1	0.99	0.99	0.99	0.99	5	5	5	5	1	1	1	1
\Lidocaine	Literature	0	0.01	0.00	0.00	0.04	3	3	1	1	1	1	1	1
\Lidocaine	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Losartan	Literature	0	0.18	0.01	0.00	0.00	2	1	0	0	1	1	1	1
\Losartan	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Losartancarboxylic acid	MASST	1	0.64	0.69	0.69	0.69	5	5	6	6	1	1	1	1
\LosartanMinus2H	MASST	1	0.00	0.55	0.67	0.68	0	7	8	9	0	1	1	1
\LosartanMinus2HPlusO	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Melitracen	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Methadone	Literature	1	0.38	0.39	0.53	0.60	7	7	6	6	1	1	1	1
\Metoprolol	Literature	1	0.93	0.97	0.98	0.98	4	4	4	4	4	4	4	3
\MetoprololMinus2HPlusO	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\NNDidesmethylvenlafaxine	Literature	0	0.14	0.26	0.83	1.00	1	1	0	1	1	1	1	1
\NNDidesmethylvenlafaxine	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Ndesmethyltramadol	Literature	0	0.14	0.24	0.00	1.00	2	2	0	1	1	1	1	1
\Ndesmethyltramadol	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Odesmethylvenlafaxine	Literature	0	0.02	0.00	0.00	0.00	3	3	0	0	1	1	1	1
\Odesmethylvenlafaxine	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Propiconazole	Literature	0	0.00	0.00	0.01	0.02	0	0	0	0	1	1	1	1
\Propiconazole	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Ritalinicacid	Literature	1	0.06	0.00	0.62	0.68	2	2	1	2	1	1	1	1
\Tapentadol	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Tebutryn	Literature	1	0.47	0.02	0.99	0.99	4	4	4	4	4	4	4	4
\Tramadol	Literature	0	0.02	0.00	0.00	0.00	3	3	0	0	1	1	1	1
\Tramadol	MASST	0	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0
\Venlafaxine	Literature	0	0.02	0.01	0.00	0.00	1	1	0	0	0	0	0	0
\Venlafaxine	MASST	1	0.00	0.00	0.84	0.80	0	0	2	3	0	0	9	9
\Verapamilmetabolite	MASST	1	0.00	0.00	0.56	0.68	0	0	5	2	0	0	1	1

S4: Threshold evaluation for all suspects



S5: ROI Artefacts and the Effect of Smoothing on MCR and MF Performance

Region of interest (ROI) binning is a critical pre-processing step in LC-HRMS data analysis (as outlined in **S2**) that significantly impacts subsequent analytical methods, such as MCR and MF. The ROI procedure involves the selection of mass spectral data within specified regions that exceed a certain threshold of interest. This process can introduce artefacts, particularly when data points are missing or inconsistent, potentially distorting the EICs of fragments.

The MCR relies on the assumption that all fragments of a component exhibit the same elution profile and hence artefacts in the EICs can make the model less robust. Discrepancies in the fragment EICs can cause the MCR algorithm to inappropriately distribute fragments across different components. In the example of Terbutryn, a 3-component MCR model resolves the pure component MS2 spectra accurately, after prior mass filtering (MF+MCR data analysis strategy). However, overfitting the model slightly by using a 5 component model leads to the discrimination of the mass fragment with mz 186.0815. We compared this situation to a 3-component model and a 5-component model fitted on data that has been smoothed with a moving average of window sizes W(2,2) and W(3,3). For a 3-component model, spectra of Terbutryn extracted from non-smoothed and smoothed data look completely alike, as can be seeen in **Figure 6A**. The situation

changes for a 5-component model, shown in **Figure 6B**, where the mass fragment with mz 186.0815 is contained in the mass spectra extracted from the smoothed data, while it has been modeled by a different component in the case of non-smoothed data.

Our investigation revealed that the correlation coefficients between mass fragment EICs mz 186.0815 and mz 187.0825 increased from 0.93 and 0.95 in the non-smoothed data to 0.97 and 0.97 after smoothing with a moving average window W(2,2), and further to 0.98 and 0.98 when a broader moving average window W(3,3) was used. The smoothing process mitigated the impact of ROI artefacts, aligning the elution profiles of the fragments more closely, as depicted by **Figure 6C**.

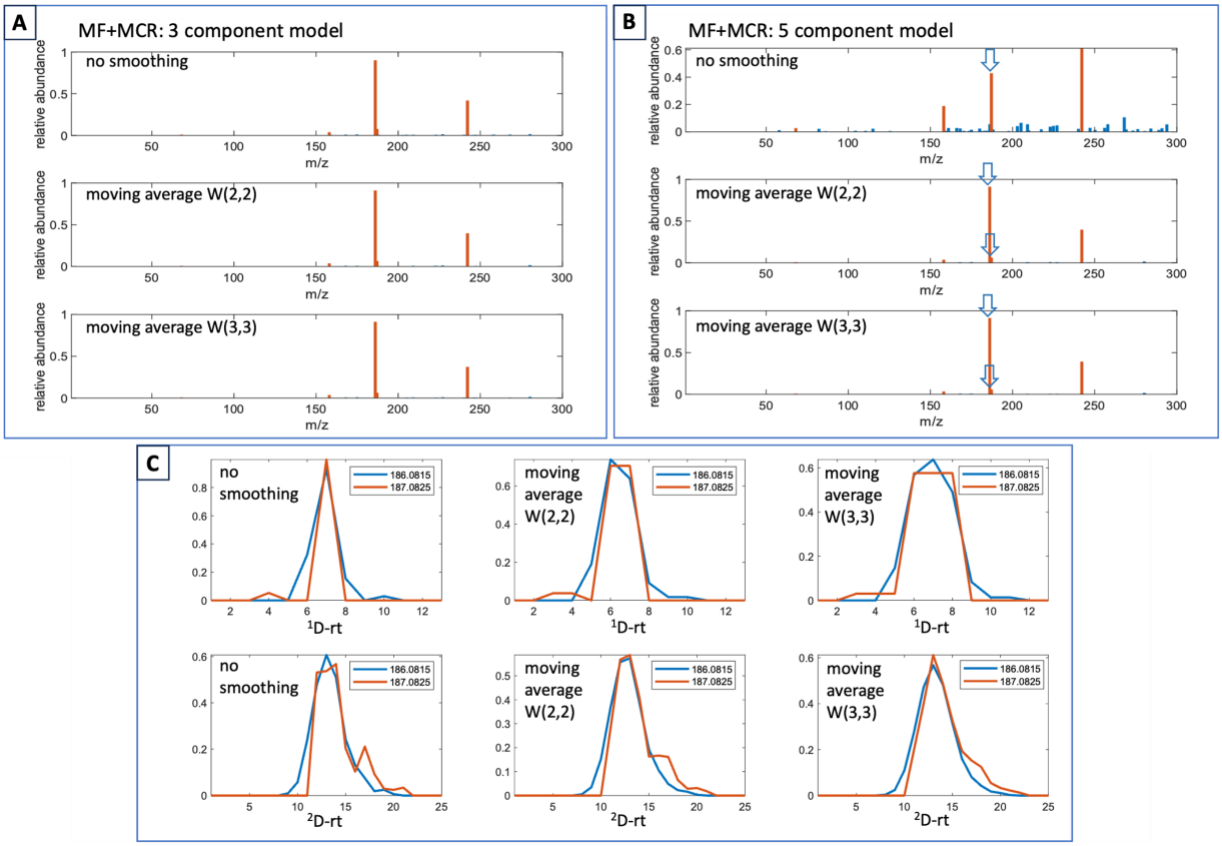


Figure 6*: Visual representation of the impact of ROI binning and smoothing on MCR model robustness.* ***A****: we observe that a 3-component MCR model yields consistent spectra for Terbutryn from both non-smoothed and smoothed data.* ***B****: A 5-component MCR model, which slightly overfits the data, incorrectly discriminates against the mass fragment mz 186.0815 in non-smoothed data but correctly includes it in smoothed data using moving average filters of different window sizes.* ***C****: The improved alignment of elution profiles due to smoothing mitigates ROI artefacts as shown for mz 186.0815 and mz 187.025.*

However, smoothing can also have a detrimental effect. Specifically, we observed that smoothing reduced the specificity in the DAS3 (MF). The MF method's performance hinges on the specificity of fragment EICs to correctly group EICs based on their similarity to a reference EIC. Smoothing these EICs can blur subtle differences that are pivotal for the accurate identification and grouping

by MF. This effect can be emphasized in **Figure 7** in which the distribution of ${}^{12}_{total}S$ is shown for the non-smoothed and smoothed EICs in the ${}^{1}t_r \times {}^{2}t_r$ frame of Terbutryn. The distributions clearly show a shift to higher score values if the EICs have been smoothed, which results in a higher number of false positive fragments passing the filter.

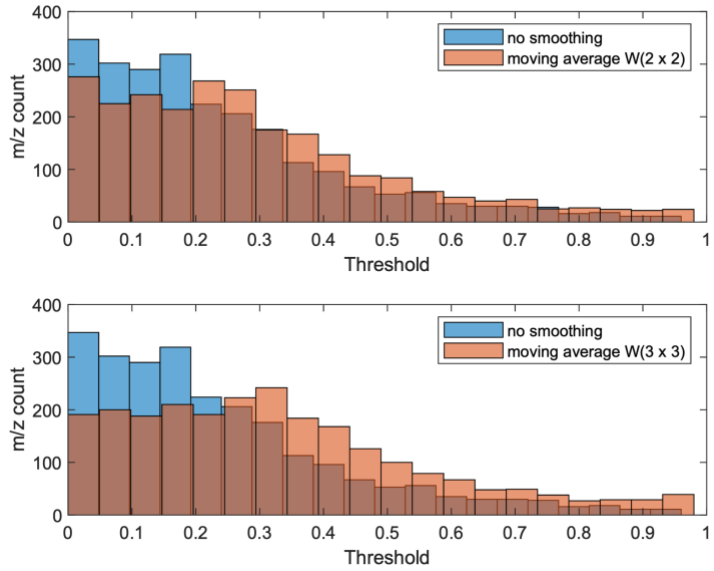


***Figure 7**: illustrates the counteractive effects of smoothing on the specificity of the Mass Filtering (MF). It is evident that post-smoothing, the score distribution shifts towards higher values, indicative of a loss in specificity and an potentially increased rate of false positives passing the MF filter.*

Given these findings, we propose a two-step process to balance the benefits and drawbacks of smoothing: Conduct Mass Filtering prior to any smoothing to preserve specificity, then apply smoothing to the selected masses. For instance, a moving average filter can be employed post-MF to enhance the robustness of the subsequent MCR analysis. This sequential approach leverages the strength of both methods - MF's specificity in selecting relevant mass traces and MCR's capability to deconvolute co-eluting peaks - while mitigating their respective weaknesses.